JDS 2021
52èmes Journées de Statistique
de la Société Française de Statistique (SFdS)

Recueil des soumissions

Table des matières

Classification topologique de variables

Rafik Abdesselam

Laboratoires ERIC - COACTIS, Université de Lyon, Lumière Lyon 2,
16, quai Claude Bernard 69365 Lyon cedex 07
rafik.abdesselam@univ-lyon2.fr

La classification d'objets (individus ou variables) est l'une des approches les plus utilisées pour explorer les données multivariées. Les stratégies de classification non supervisée les plus courantes sont la Classification Ascendant Hiérarchique (CAH) et la classification non hiérarchique des centres mobiles (k-means), utilisées pour identifier des groupes d'objets similaires dans un ensemble de données pour le partitionner en groupes homogènes. La Classification Topologique de Variables proposée, notée CTV, est basée sur la notion de graphes de voisinage. Certaines variables sont plus ou moins corrélées ou liées selon le type quantitatif et/ou qualitatif des variables. Cette approche topologique d'analyse des données peut alors être utile pour la réduction de dimension et la sélection de variables. Il s'agit d'une analyse hiérarchique topologique de regroupement d'un ensemble de variables de tout type. Des exemples sur données réelles illustrent cette approche dont les résultats sont comparés à ceux d'autres approches de classification de variables.

Mots-clés. Classification, mesure de proximité, matrice d'adjacence, graphe de voisinage, variables mixtes.

Abstract. The clustering of objects (individuals or variables) is one of the most important approaches to exploring multivariate data. The most common unsupervised clustering strategies are Hierarchical Ascending Clustering (HAC) and k-means partitioning used to identify groups of similar objects in a dataset to divide it into homogeneous groups. The proposed Topological Clustering of Variables, called TCV, studies an homogeneous set of variables defined on the same set of individuals, based on the notion of neighborhood graphs. Some the variables are more-or-less correlated or linked according to the type quantitative or qualitative of the variables. This topological approach of data analysis can then be useful for dimension reduction and variable selection. Its a topological hierarchical clustering analysis of a set of variables of any type. Examples on real data illustrate this approach whose results are compared with those of other variable classification approaches.

Keywords. Clustering, proximity measure, adjacency matrix, neighborhood graph, mixed variables.

1 Introduction

La présente étude propose une approche topologique pour la classification de variables CTV, sans aucune restriction sur le type des variables, quantitatives, qualitatives ou un

mélange des deux. Outre les méthodes classiques et bien connues consacrées à la classification d'objets, il existe des approches spécifiquement dédiées à la classification de variables telles que la procédure de classification Varclus implémentée sous SAS (Varclus procédure (2018)), l'approche ClustOfVar (Chavent et al.(2012)), l'approche CLV (Vigneau et Qannari(2003)) pour regrouper les variables autour des composantes latentes et l'approche Clustatis ((Llobell and Qannari (2019)), mais à notre connaissance, aucune approche n'est proposée dans un contexte topologique.

Le processus consiste à créer des groupes de variables similaires, c'est-à-dire portant la même dimension d'information, les variables d'un même groupe sont les plus liées les unes aux autres tandis que les variables des différents groupes sont les plus orthogonales que possible. L'objectif ici est d'analyser les structures sous-jacentes des données, de constituer une synthèse des informations portées par les variables en vue d'une réduction de dimension et/ou d'une sélection de variables explicatives. Contrairement à la classification d'individus qui se fait généralement à partir d'un seul ensemble de variables homogènes relatives à un même thème, la classification de variables permet de traiter plusieurs ensembles de variables homogènes de plusieurs thèmes différents.

Les mesures de similarité jouent un rôle important dans de nombreux domaines de l'analyse des données, les résultats de toute opération de structuration, de regroupement ou de classification d'objets dépendent fortement de la mesure de proximité choisie. Des études factorielles topologiques ont été proposées notamment dans le contexte de l'analyse discriminante (Abdesselam (2019)), des analyses des correspondances simples et multiples (Abdesselam (2019)) et de l'analyse en composantes principales (Abdesselam (2020)), mais aucune sur la classification topologique de variables.

L'approche CTV est illustrée par des exemples sur données réelles dont les résultats sont comparés avec ceux de différentes approches connues de classification de variables.

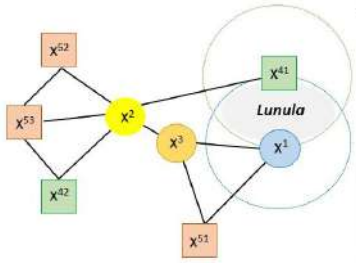
2 Contexte topologique

L'analyse topologique des données est une approche basée sur le concept de graphe de voisinage. Etant données une mesure de proximité et une structure topologique choisie, on peut faire correspondre un graphe topologique induit sur l'ensemble des objets.

Soit $E = \{x^1, \cdots, x^j, \cdots x^p, y^{11}, \cdots, y^{1m_1}, \cdots, y^{q1}, \cdots, y^{qm_q}\}$ un ensemble de variables mixtes, constitué de p variables quantitatives $\{x^j\}_{j=1,p}$ et de q variables qualitatives $\{y^k\}_{k=1,q}$, où $m = \sum_{k=1}^{q} m_k$ est le nombre total de modalités et m_k désigne le nombre de modalités de la variablee y^k. A l'aide d'une mesure de proximité notée u, on peut définir une relation de voisinage notée V_u qui est une relation binaire sur $E \times E$ où les sommets sont les variables et les arêtes sont définies par une relation de voisinage. Il existe de nombreuses définitions pour construire cette relation binaire de voisinage, par exemple, l'Arbre de Longueur Minimale (ALM), le Graphe de Gabriel (GG), ou encore le Graphe des Voisins Relatifs (GVR) (Toussaint (1980)). On peut établir la matrice dite

d'adjacence V_u associée à u, d'ordre $p+m$, binaire et symétrique, où toutes les paires de variables voisines $(x^k, x^l)_{k,l=1,p}$ dans E, satisfont la propriété GVR suivante :

$$V_u(x^k, x^l) = \begin{cases} 1 & \text{si } u(x^k, x^l) \leq \max[u(x^k, x^t), u(x^t, x^l)]; \forall x^k,\ x^l,\ x^t \in E, \quad x^t \neq x^k \text{ et } x^t \neq x^l \\ 0 & \text{sinon.} \end{cases}$$



V_u	x^1	x^2	x^3	x^{41}	x^{42}	x^{51}	x^{52}	x^{53}
x^1	1	0	0	1	0	0	0	0
x^2	0	1	1	0	0	0	0	0
x^3	0	1	1	0	0	0	0	0
x^{41}	1	0	0	1	0	0	0	0
x^{42}	0	0	0	0	1	0	0	1
x^{51}	0	0	0	0	0	1	0	0
x^{52}	0	0	0	0	0	0	1	1
x^{53}	0	0	0	0	1	0	1	1

Figure 1: GVR d'un ensemble de variables mixtes - Matrice d'adjacence associée

La mesure de proximité u génère une structure topologique sur les variables dans E qui est totalement décrite par la matrice d'adjacence V_u.

La Figure 1 illustre un exemple d'un ensemble de huit objets, trois variables quantitatives $\{x^1, x^2, x^3\}$ et cinq variables indicatrices $\{x^{41}, x^{42}, x^{51}, x^{52}, x^{53}\}$ de deux variables qualitatives $\{x^4, x^5\}$, qui vérifient la propriété des GVR. Par exemple, pour la première variable quantitative x^1 et la première variable indicatrice x^{41}, $V_u(x^1, x^{41}) = 1$, cela signifie sur le plan géométrique que l'hyper-Lunule, intersection des deux hypersphères centrées sur les deux points-variables x^1 et x^{41}, est vide.

2.1 Matrices d'adjacence de référence

Trois approches topologiques selon le type de variables considérées, quantitatives ou qualitatives ou un mélange des deux, sont proposées. Elles sont basées sur la matrice d'adjacence V_{u^*} associée à la mesure de proximité u^* inconnue dite de référence. L'objectif ici est de décrire la structure topologique de corrélation ou de dépendance entre les variables considérées.

- Soit $\{x^j\}_{j=1,p}$ un ensemble de p variables quantitatives observées sur n individus-objets. Dans ce cas, la matrice d'adjacence $V_{u_\star}$ est définie à partir du test t de Student du coefficient de corrélation linéaire ρ de Bravais-Pearson :

$$V_{u_\star}(x^k, x^l) = \begin{cases} 1 & \text{si p-valeur } = P[\mid T_{n-2} \mid > \text{t-valeur}] \leq \alpha \ ; \ \forall k, l = 1, p \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

où la p-valeur désigne le seuil de signification α du test bilatéral d'hypothèses nulle et alternative, $H_0 : \rho(x^k, x^l) = 0$ vs. $H_1 : \rho(x^k, x^l) \neq 0$ et T_{n-2} la statistique de test de Student à $\nu = n - 2$ degrés de liberté.

• Soit $\{y^k\}_{k=1,q}$ un ensemble de q variables qualitatives observées sur $n = \sum_{k=1}^{q} n_k$ individus selon m_k modalités et qui totalisent $m = \sum_{k=1}^{q} m_k$ modalités. La matrice d'adjacence $V_{u_\star}$ est établie à partir du tableau de Burt et de l'écart à l'indépendance.

$$V_{u_\star}(y^{kr}, y^{ls}) = \begin{cases} 1 & \text{si } \frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}} \geq \frac{\mathcal{B}_{kr\,..}}{nq^2} \quad ; \quad \forall k, l = 1, q \ ; \ r = 1, m_k \ \ et \ \ s = 1, m_l \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

$\mathcal{B}_{kr\,ls} = \Sigma_{i=1}^{n} y_i^{kr} y_i^{ls}$, élément de la matrice de Burt qui correspond au nombre d'individus qui possède la modalité r de y^k et la modalité s de y^l,
$\mathcal{B}_{kr\,..} = \Sigma_{l=1}^{q} \Sigma_{s=1}^{m_s} b_{kr\,ls}$ désigne la marge ligne de la modalité r de y^k,
$\frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}}$ désigne le profil-ligne de la modalité r de y^k,
$\frac{\mathcal{B}_{kr\,..}}{nq^2}$ est le profil-moyen de la modalité r de y^k, nq^2 est le nombre total.

• Soit $\{x^1, \cdots, x^j, \cdots, x^p, y^1, \cdots, y^k, \cdots, y^q\}$ un ensemble de données mixtes constitué de p variables quantitatives et q variables qualitatives, observées sur n individus.

Le traitement simultané de données mixtes ne peut pas être réalisé directement par les méthodes conventionnelles d'analyse des données. Ainsi, dans un premier temps, les données qualitatives sont transformées en données quantitatives, en utilisant l'analyse en composantes principales mixte (Abdesselam (2008)), basée sur la maximisation du critère mixte, proposé en termes de carrés de corrélation par Saporta (Saporta 1990)) et géométriquement en termes de cosinus carrés des angles par Escofier (Escofier (1979)). On peut également effectuer une analyse factorielle des données mixtes développée par Pagès (Pagès (2004)). La matrice d'adjacence $V_{u_\star}$ est ensuite établie à partir de la matrice de corrélation de l'ensemble des variables, quantitatives et qualitatives transformées, selon l'expression (1).

3 Classification topologique

Afin de décrire les similitudes entre les variables et les regrouper en groupes homogènes, on applique la technique dite du thémascope (Lebart (1989)), qui consiste en un enchaînement méthodologique d'une méthode de classification sur les composantes principales d'une méthode d'analyse factorielle, en l'occurence ici une analyse factorielle topologique suivie d'une classification ascendante hiérarchique (HAC) selon le critère de Ward (Ward (1963)).

L'analyse factorielle topologique sous-jacente est soit une ACPT si les variables sont quantitatives (Abdesselam (2020)), soit une ACMT si les variables sont qualitatives (Abdesselam (2019)) ou une ACPMT si les variables sont mixtes (Abdesselam (2008, 2020)).

Pour la méthode d'analyse factorielle topologique sous-jacente, on effectue la technique du Multidimensional Scaling (MDS), à savoir l'analyse factorielle sur un tableau de similarité (Caillez and Pagès (1976)), la matrice d'adjacence de référence $V_{u_\star}$ associée à la mesure de proximité u^*, la mesure la plus adaptée aux données considérées.

L'approche hiérarchique CTV et son dendrogramme sont facilement programmables à partir des procédures ACP et CAH des logiciels SPAD, SAS ou R.

Dans le cas de la CTV sur variables quantitatives, on considère que deux variables positivement ou négativement corrélées sont liées, on prend donc en compte le signe de la corrélation entre variables. Il est à noter que la procédure Varclus de SAS inclut également cette option et que c'est une méthode de classification descendante hiérarchique (CDH).

4 Exemple illustratif - Cas de variables quantitatives

Les données utilisées (Govaert (2003)) concernent 38 marques françaises d'eau en bouteille décrites par 8 variables homogènes représentant les teneurs en ions (mg/l) affichées sur les étiquettes des bouteilles. L'objectif ici est de donner une partition en classes de teneurs en ions des marques de bouteilles d'eau.

Table 1: Matrice des corrélations (p-valeurs) - Matrice d'adjacence

	CA	MG	NA	K	SULF	NO3	HCO3	CL
CA	1.0000							
MG	0.6672 (< .0001)	1.0000						
NA	0.0042 (0.9757)	0.5649 (< .0001)	1.0000					
K	0.1072 (0.4358)	0.6703 (< .0001)	0.8817 (< .0001)	1.0000				
SULF	0.8997 (< .0001)	0.5629 (< .0001)	-0.0957 (0.4872)	-0.0546 (0.6923)	1.0000			
NO3	-0.0473 (0.7317)	-0.1756 (0.1998)	-0.0830 (0.5469)	-0.1529 (0.2650)	-0.1288 (0.3486)	1.0000		
HCO3	0.1491 (0.2774)	0.6583 (< .0001)	0.9474 (< .0001)	0.8866 (< .0001)	-0.0573 (0.6776)	-0.0541 (0.6947)	1.0000	
CL	0.0578 (0.6749)	0.52094 (< .0001)	0.5646 (< .0001)	0.7187 (< .0001)	-0.0276 (0.8406)	-0.1053 (0.4443)	0.4794 (0.0002)	1.0000

$$V_{u^*} = \begin{pmatrix} 1 & & & & & & & \\ 1 & 1 & & & & & & \\ 0 & 1 & 1 & & & & & \\ 0 & 1 & 1 & 1 & & & & \\ 1 & 1 & 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

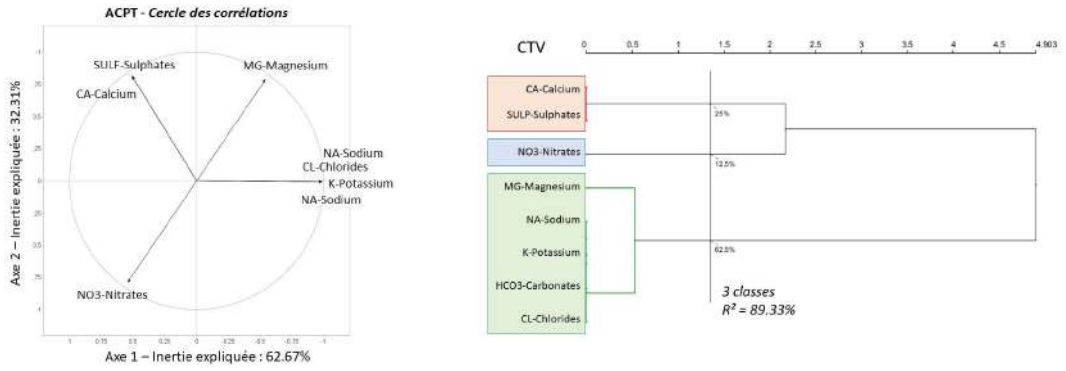


Figure 2: ACPT & CTV - Dendrogramme des teneurs en ions des marques d'eau

Le Tableau 1 présente les matrices des corrélations et d'adjacence $V_{u_\star}$ associée à la mesure de proximité $u_\star$ la plus adaptée aux données considérées, construite à partir de l'expression (1). La figure 2 illustre la représentation des variables sur le premier plan

principal de l'Analyse en Composantes Principales Topologique (ACPT). Le cercle de corrélation sur les deux premiers facteurs TPCA donne un aperçu des groupes de variables corrélées et non corrélées, une CAH selon le critère de Ward est ensuite appliquée sur les composantes principales de l'ACPT. Le dendrogramme issu de la CTV permet de visualiser et d'identifier la structure topologique des variables. Les indices d'agrégation de la CTV suggèrent une partition des 8 variables en 3 classes. On voit que les variables Calcium et Sulfates sont positivement corrélées ainsi que les variables Chlorures, Carbonates, Potassium, Sodium et Magnésium. La variable Nitrates qui constitue à elle seule la deuxième classe, est corrélée négativement avec les deux autres classes de variables.

A titre de comparaison, la Figure 3 illustre les arbres hiérarchiques des approches Varclus de SAS, ClusOfVar, CLV et Clustatis. Pour une partition en 3 classes, les constitutions des classes sont les mêmes sauf pour l'approche Varclus.

L'approche CTV en 3 classes présente un pourcentage de la variance totale expliquée ($R^2 = 89.33\%$) bien plus grand que ceux des quatre autres approches, les classes de la CTV sont ainsi beaucoup plus homogènes.

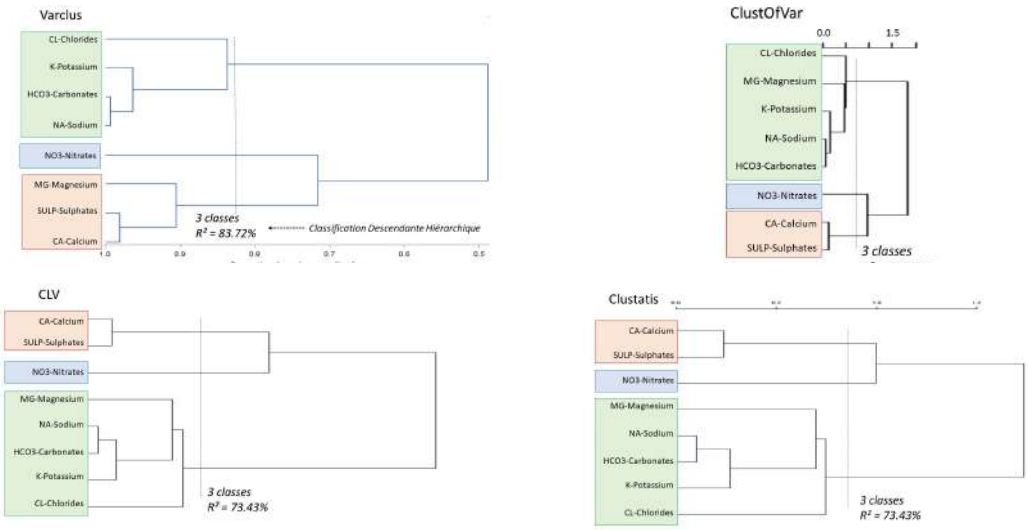


Figure 3: Dendrogrammes : Varclus, ClustOfVar, CLV et Clustatis

5 Conclusion & Perspective

Une nouvelle approche de classification de variables dans un contexte topologique est proposée, elle vient ainsi enrichir les méthodes conventionnelles de classification de données quantitatives, qualitatives ou encore mixtes. Il serait intéressant d'étendre cette approche topologique à d'autres méthodes d'analyse de données, notamment dans le cadre de l'analyse des données évolutives.

Bibliographie

[1] Abdesselam, R. (2020), A Topological Principal Component Analysis. 6th Stochastic Modeling Techniques and Data Analysis International Conference, SMTDA-2020, 2-5 June 2020, Barcelona, Spain. Virual-Online Conference.
[2] Abdesselam, R. (2019), A Topological Multiple Correspondence Analysis. *Journal of Mathematics and Statistical Science*, USA, 5, 8, 175-–192.
[3] Abdesselam, R. (2019), A Topological Discriminant Analysis. *In book Chapter, Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics*, ISTE Science Publishing LTD, Wiley, 3, 167–178.
[4] Abdesselam, R. (2008), Analyse en Composantes Principales Mixte. Classification : points de vue croisés, RNTI-C-2, *Revue des Nouvelles Technologies de l'Information* RNTI, Cépaduès Editions, 31-41, 2008.
[5] Batagelj, V., Bren, M. (1995), Comparing resemblance measures. *In Journal of classification*, 12, 73–90.
[6] Chavent M., Kuentz V., Liquet B., Saracco J., (2012), ClustOfVar: An R Package forthe Clustering of Variables, *Journal of Statistical Software*. Vol. 50, 1-16.
[7] Escofier, B. (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahier de l'analyse des données*, Vol.4(2), 137-146.
[8] Govaert, G., (2003) Analyse des données. *Hermes*, 19–42.
[9] Lesot, M-J., Rifqi, M. and Benhadda, H., (2009) Similarity measures for binary and numerical data: a survey. *In IJKESDP*, 1, 1, 63-84.
[10] Llobell, F. and Qannari, E.M., (2019) Clustering datasets by means of Clustatis with identification of atypical datasets. Application to sensometrics. *Food Quality and Preference*, Elsevier, 75, 97–104.
[11] Pagès,J. (2004), Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 52(4), 93–111.
[12] SAS Institute Inc. SAS/STAT Software, Version 9.3 user's guide, The Varclus Procedure, URL http://support.sas.com/documentation/onlinedoc/stat/930/varclus.pdf.
[13] Saporta, G. (1990), Simultaneous treatment of quantitative and qualitative data. *In Attidela XXXV Riunone scientifica; Società Italiana di Statistica*, 63–72.
[14] Toussaint, G. T. (1980), The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268.
[15] Vigneau, E., Qannari, E.M. (2003), Classification of variables around latent components. *Communications in statistics Simulation and Computation*, 32(4), 1131-1150.
[16] Ward, J-R. (1963) Hierarchical grouping to optimize an objective function. *In Journal of the American statistical association JSTOR*, 58, 301, 236–244.
[17] Zighed, D., Abdesselam, R., and Hadgu, A. (2012), Topological comparisons of proximity measures. *16th PAKDD 2012 Conference*, Part I, LNAI 7301, Springer, 379–391.

Tester la symétrie des marginales des distributions pour des images digitales de bruit grâce aux périmètres d'ensembles de niveaux

Mariem Abaach[1], Hermine Biermé[2] & Elena Di Bernardino[3]

[1] *MAP5 UMR CNRS 8145, Université de Paris*
45 rue des Saints-Pères, 75006 Paris, France. mariem.abaach@u-paris.fr.
[2] *LMA UMR CNRS 7348, Université de Poitiers,*
11 bd Marie et Pierre Curie, 86962 Chasseneuil, France.
hermine.bierme@math.univ-poitiers.fr.
[3] *Laboratoire J.A. Dieudonné, UMR CNRS 7351,*
Université Côte d'Azur, Nice, France. Elena.Di_bernardino@unice.fr.

Résumé. Notre étude porte sur des images digitales dont la valeur des pixels est supposée donnée par une suite de variables aléatoires indépendantes et identiquement distribuées dans une fenêtre d'observation fixée. On commence par proposer un estimateur non biaisé du périmètre qui ne prend pas en considération les effets de bord induits par la fenêtre d'observation. L'étude du premier et second moment de cet estimateur nous permet d'établir un résultat de normalité asymptotique auto-normalisé muni d'une matrice de covariance explicite et accessible empiriquement. Ce Théorème Central Limite nous permet de construire un test statistique consistant et accessible afin de tester la symétrie des distributions marginales. Dans un second temps, on s'intéresse au comportement asymptotique du périmètre dans un régime limite particulier.
Mots-clés. Images binaires, Franchissements, Ensembles de niveaux, Test de symétrie, Procédure de seuillage.

Abstract. In this paper we consider digital images for which the pixels values are given by a sequence of independent and identically distributed variables within an observation window. We proceed to the construction of an unbiased estimator for the perimeter without border effects. The study of the first and second moments of the perimeter allows to prove auto-normalised asymptotic normality results with an explicit covariance matrix consistently estimated. Theses Central Limit Theorems permit to built a consistent and empirical accessible test statistic to test the symmetry of the marginal distribution. Finally the asymptotic perimeter behaviour in large threshold limit regime is also explored. Several numerical studies are provided to illustrate the proposed testing procedures.
Keywords. Binary image, Crossings, Excursion sets, Test of symmetry, Threshold procedure

1 Introduction

La modélisation stochastique des images par des champs aléatoires permet la mise en place d'un cadre statistique intéressant pour les différentes problématiques de traitements d'images telles que le débruitage d'images, la reconnaissance de formes, la segmentation ou encore la classification. En particulier, l'étude des attributs géométriques des objets a suscité un grand intérêt ces dernières années. Moralement, ses caractéristiques peuvent s'interpréter comme l'aire, le périmètre et la caractéristique d'Euler (*i.e.* le nombre de composantes connexes - le nombre de trous) d'une image en noir et blanc obtenue en seuillant l'image en niveau de gris à un niveau fixé (voir [5] pour une introduction formelle de ces objets). Ces nouvelles images en noir et blanc peuvent représenter les ensembles d'excursions d'un champ aléatoire sous-jacent. Ses caractéristiques géométriques, aussi appelées les courbures de Lipschitz-Killing (LK), sont des descripteurs de forme robustes avec un large domaine d'application dont la médecine (*e.g.* l'étude de mammographies digitales synthétiques 2D [3, 4]).
D'importants résultats ont déjà été démontrés dans le cadre de champs aléatoires lisses supposés stationnaires en particulier pour le champ aléatoire Gaussien (voir [1]). Dans ce cadre, la moyenne théorique ainsi que la variance des courbures LK peuvent être calculées explicitement en fonction des paramètres du champ et peuvent être estimées de manière consistante à partir des images. De plus, des résultats de type Théorème Central Limite ont été démontrés ce qui a permis d'établir de nombreux tests statistiques.
Dans cet article, on se place dans un cadre discret où l'on considère des images digitales pour lequel la notion de bruit blanc est bien définie. Plus précisément, on suppose que la valeur des pixels $(X_{i,j})_{1\leq i,j\leq m}$ est donnée par une suite de variables aléatoires indépendantes et identiquement distribuées dans une fenêtre d'observation S, comme par exemple dans l'étude du résidu d'une procédure de débruitage ou d'une régression linéaire ou encore la différence de deux images (*e.g* dans le cadre de l'étude de l'activité cérébrale). Ainsi, on souhaite tester l'hypothèse naturelle de symétrie qui est de savoir si les $X_{i,j}$ sont tirés suivant une loi symétrique. L'hypothèse de symétrie des marginales est appelé *hypothèse nulle* H_0. Être capable de tester formellement l'hypothèse de symétrie est un sujet important qui trouve des applications dans nombreux domaines, puisque cette hypothèse contient des nombreuses informations sur le champ sous-jacent.

2 Cadre mathématique

Soit m un entier relatif avec $m \geq 2$. On se place sur le carré unité $S = [0,1]^2$ et on divise S en m^2 carrés de cotés de longueur $1/m$, *i.e*,

$$C_{i,j}^{(m)} := \left[\frac{i-1}{m}, \frac{i}{m}\right] \times \left[\frac{j-1}{m}, \frac{j}{m}\right], \quad \text{pour } i,j \in \{1,\dots,m\}.$$

Chaque $C_{i,j}^m$ est appelé *cellule* et $S = \bigcup_{1\leq i,j\leq m} C_{i,j}^m$. Soit $(X_{i,j})_{1\leq i,j\leq m}$ une suite de variables aléatoires définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Les variables $(X_{i,j})_{1\leq i,j\leq m}$ sont supposées indépendantes et identiquement distribuées. La valeur de chaque pixel $X_{i,j}$ est associé à une cellule $C_{i,j}$. Dans la Figure 1, $X_{i,j} \sim \mathcal{U}(0,1)$ et $m = 20$.

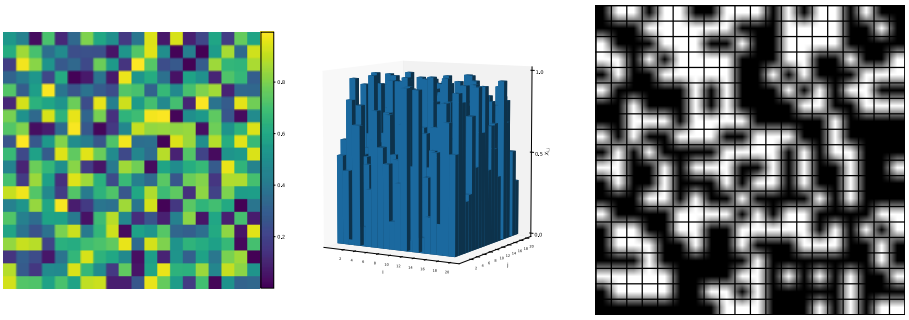


Figure 1: De gauche à droite: Image de taille (20×20) réalisation d'un bruit blanc Uniforme (représentation en 2D et 3D) et l'image binaire obtenue pour une valeur de seuil $t = 0.5$.

Soit $t \in \mathbb{R}$ une valeur de seuil fixée, pour former l'image binaire associée, on introduit une matrice aléatoire $Z(t)$ de taille $m \times m$:

$$Z_{i,j}^{(m)}(t) := \mathbb{1}_{\{X_{i,j}\geq t\}}, \text{ pour } i,j \in \{1,\ldots,m\}.$$

Alors, $Z_{i,j}(t)$ suit une loi Binomiale de paramètre $(1, p_t)$, avec

$$p_t := \mathbb{P}(Z_{i,j}(t) = 1) = \mathbb{P}(X_{i,j} \geq t) = 1 - F(t^-),$$

où F est la fonction de répartition associée à $X_{i,j}$. Ainsi, à chaque cellule $C_{i,j}^m$, on associe une couleur noire ou blanche suivant que la cellule associée $Z_{i,j}(t)$ est égale à 0 ou 1.

Périmètre d'une image binaire Étant donné l'image binaire $Z = Z(t)$ obtenue pour un seuil $t \in \mathbb{R}$. Comme présenté dans l'article [2], le périmètre de l'image Z peut être défini comme étant égale à la somme de toutes les arêtes des cellules $Z_{i,j}^m$ qui appartiennent à la frontière de la partie noire de l'image binaire Z. Pour cela on peut introduire les quantités:

$$f_1^{(t)}(l,k) = \mathbb{1}_{\{(Z_{l,k-1}(t)=0 \cap Z_{l,k}(t)=1) \cup (Z_{l,k-1}(t)=1 \cap Z_{l,k}(t)=0)\}}$$

pour compter les contributions horizontales et

$$f_2^{(t)}(k,l) = \mathbb{1}_{\{(Z_{k-1,l}(t)=0 \cap Z_{k,l}(t)=1) \cup (Z_{k-1,l}(t)=1 \cap Z_{k,l}(t)=0)\}}$$

pour compter les contributions verticales.

Definition 2.1. *On note* $\mathcal{P}_m^{(1)}(t) = \sum_{l=1}^m \sum_{k=2}^m f_1^{(t)}(l,k)$ *la somme de toutes contributions horizontales et* $\mathcal{P}_m^{(2)}(t) = \sum_{l=1}^m \sum_{k=2}^m f_2^{(t)}(k,l)$ *la somme des contributions verticales, alors, la valeur du périmètre et du périmètre normalisé est donnée par les formules suivantes*

$$\mathcal{P}_m(t) := \mathcal{P}_m^{(1)}(t) + \mathcal{P}_m^{(2)}(t), \qquad \check{\mathcal{P}}_m(t) := \frac{1}{m^2}(\mathcal{P}_m^{(1)}(t) + \mathcal{P}_m^{(2)}(t)).$$

3 Statistiques du périmètre d'une image binaire

Le premier moment du périmètre est alors donné par :

Proposition 3.1. *L'espérance du périmètre normalisé dans la Définition (2.1) est donnée par*

$$\mathbb{E}(\check{\mathcal{P}}_m(t)) = 4p_t(1-p_t)\left(1-\frac{1}{m}\right) := \mu_{\check{\mathcal{P}}}(p_t, m). \tag{1}$$

On remarque alors que si $X_{i,j}$ est tiré suivant une loi continue d'axe de symétrie $\theta \in \mathbb{R}$, *i.e* $p_{\theta-t} = 1 - p_{\theta+t}$, $\forall t \in \mathbb{R}$. Alors, $\mu_{\check{\mathcal{P}}}(p_{\theta-t}, m) = \mu_{\check{\mathcal{P}}}(p_{\theta+t}, m)$ et donc, $t \mapsto \mu_{\check{\mathcal{P}}}(p_t, m)$ admet un axe de symétrie en θ. Ceci implique que le ratio $(\check{\mathcal{P}}_m(\theta-t))/(\check{\mathcal{P}}_m(\theta+t))$ devrait être distribué aux alentours de 1 dans le cas symétrique.

L'étude de la covariance du périmètre nous permet d'établir un TCL.

Theorem 3.2. *Soit* r *un entier positif,* $m \geq 2$ *et* $t_1, \ldots, t_r \in \mathbb{R}$*, alors,*

$$m\left(\begin{pmatrix}\check{\mathcal{P}}_m(t_1)\\ \check{\mathcal{P}}_m(t_2)\\ \vdots \\ \check{\mathcal{P}}_m(t_r)\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\check{\mathcal{P}}_m(t_1))\\ \mathbb{E}(\check{\mathcal{P}}_m(t_2))\\ \vdots \\ \mathbb{E}(\check{\mathcal{P}}_m(t_r))\end{pmatrix}\right) \xrightarrow[m\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_r^\star),$$

avec $\xrightarrow{\mathcal{L}}$ *représentant la convergence en loi et* $\mathcal{N}(0, \Sigma_r^\star)$ *la distribution Gaussienne en dimension* r *centré de matrice de covariance* $\Sigma_r^\star$ *donnée par*

$$\Sigma_r^\star(i,j) := 4p_{\max(t_i,t_j)}(1-p_{\min(t_i,t_j)})\Big(4-14p_{\min(t_i,t_j)}(1-p_{\max(t_i,t_j)})+6(p_{\min(t_i,t_j)}-p_{\max(t_i,t_j)})\Big).$$

La dernière brique nécessaire pour construire un test accessible est l'estimation de p_t. Elle est donnée par la définition suivante.

Definition 3.3. *Étant donné* $t \in \mathbb{R}$*, on partitionne l'image* Z *en* $m^2/4$ *sous images* Z_2^i *de taille* (2×2)*. On dénote par* $\check{\mathcal{P}}_2^i(t)$ *la valeur du périmètre normalisé de chaque sous image. Soit* $\overline{S}_m(t) := \frac{4}{m^2}\sum_i \check{\mathcal{P}}_2^i(t)$ *et* $g : [0,1] \to \mathbb{R}$ *la fonction continue, définie par*

$$x \mapsto g(x) = \begin{cases} \frac{1}{2}\left(1 - \sqrt{1-2x}\right) & si\ x < \frac{1}{2}, \\ \frac{1}{2} & sinon. \end{cases}$$

Alors, pour $t > \theta$ avec θ la valeur médiane de la distribution, on définit un estimateur de $p(t)$ basée sur le périmètre par

$$\widehat{p}_m^{\mathcal{P}}(t) := g(\overline{S}_m(t)).$$

On arrive à montrer que l'estimateur $\widehat{p}_m^{\mathcal{P}}(t)$ est asymptotiquement normal.

4 Test de symétrie basé sur le périmètre

On définit l'hypthèse nulle $H_0(t)$ pour $t \in \mathbb{R}$ tel que, $0 < p_{t+\theta} < \frac{1}{2}$,

$$H_0(t) \quad : \quad p_{\theta-t} = 1 - p_{\theta+t},$$

avec θ la valeur médiane de la distribution.
Soit $t \in \mathbb{R}$ tel que $0 < p_{t+\theta} < \frac{1}{2}$ et notons $R_{m,\theta}(t) := \dfrac{\check{\mathcal{P}}_m(\theta - t)}{\check{\mathcal{P}}_m(\theta + t)}$, alors sous $H_0(t)$,

$$m\left(R_{m,\theta}(t) - 1\right) \xrightarrow[m\to\infty]{d,H_0} \mathcal{N}\left(0, \sigma^2(\theta + t)\right),$$

avec

$$\sigma^2(\theta + t) = \frac{(2p_{\theta+t} - 1)(3p_{\theta+t} - 2)}{p_{\theta+t}(1 - p_{\theta+t})^2}.$$

Soit $\alpha \in (0, 1)$ et $q_{1-\alpha/2}$ tel que $\mathbb{P}(\mathcal{N}(0, 1) \leq q_{1-\alpha/2}) = 1 - \alpha/2$. On peut donc définir un test statistique de niveau α par

$$\phi_m^{\mathcal{P}}(\sigma) = \mathbb{1}_{\left\{|\frac{m}{\sigma(\theta+t)}(R_{m,\theta}(t) - 1)| \geq q_{1-\alpha/2}\right\}},$$

et un test accessible de niveau α en utilisant $\widehat{p}_m^{\mathcal{P}}(t)$ dans la Définition (3.3)

$$\phi_m^{\mathcal{P}}(\widehat{\sigma}^{\mathcal{P}}) = \mathbb{1}_{\left\{|\frac{m}{\widehat{\sigma}_m^{\mathcal{P}}(\theta+t)}(R_{m,\theta}(t) - 1)| \geq q_{1-\alpha/2}\right\}}. \tag{2}$$

Dans la Figure 2 nous presentons des résultats numériques sous H_0 et H_1.

Idées en perspectives En explorant l'équivalence entre la variance du périmètre et sa moyenne dans un régime limite particulier, on arrive à construire un test pour larges niveaux qui permet de se défaire de l'estimation de la variance dans le test $\phi_m^{\mathcal{P}}(\widehat{\sigma}^{\mathcal{P}})$ (voir Équation (2)). Nous sommes actuellement en train d'étudier le comportement du premier et second moment du périmètre dans un champ Gaussien corrélé.

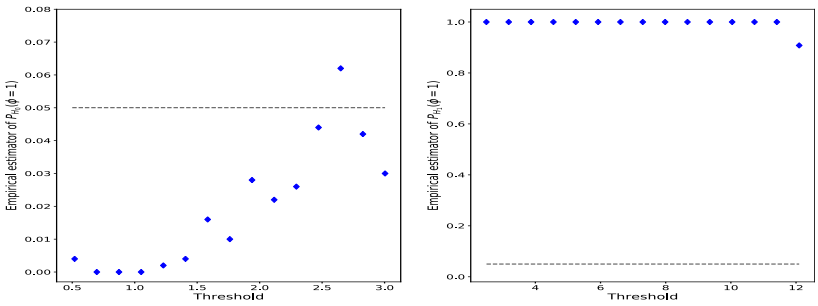


Figure 2: **Sous l'hypothèse H_0 et H_1.** On estime, pour différents seuils t, pour 500 simulations Montecarlo et pour $m = 512$, la valeur empirique moyenne de $\mathbb{P}_{H_0(t)}(\phi_m^{\mathcal{P}} = 1)$ pour une distribution de marginale Gaussienne $\mathcal{N}(0, 1)$ (à gauche) et la valeur moyenne de $\mathbb{P}_{H_1(t)}(\phi_m^{\mathcal{P}} = 1)$ pour une distribution exponentielle $\mathcal{E}(1)$ (à droite), pour $\phi_m^{\mathcal{P}}(\widehat{\sigma}^{\mathcal{P}})$ dans l'Équation (2).

References

[1] R. J. Adler. *The Geometry of Random Field*. John Wiley & Sons, 1981.

[2] H. Biermé and A. Desolneux. The effect of discretization on the mean geometry of a 2D random field. Preprint, 2020.

[3] H. Biermé, E. Di Bernardino, C. Duval, and A. Estrade. Lipschitz-Killing curvatures of excursion sets for two-dimensional random fields. *Electronic Journal of Statistics*, 13(1):536–581, 2019.

[4] E. Di Bernardino and C. Duval. Statistics for Gaussian random fields with unknown location and scale using Lipschitz-Killing curvatures. *Scandinavian Journal of Statistics*, n/a(n/a):1–42, 2020.

[5] C. Thäle. 50 years sets with positive reach - a survey. *Surveys in Mathematics and its Applications*, 3:123–165, 2008.

Discrimination entre et dans les classes (semi)continues des modèles Tweedie et géométriques Tweedie

Rahma Abid [1] & Célestin C. Kokonendji [2]

[1] *Université Paris-Dauphine Tunis. rahma.abid@dauphine.tn*
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon. celestin.kokonendji@univ-fcomte.fr*

Résumé. Dans les modèles Tweedie et géométriques Tweedie, le paramètre de puissance commun $p \notin (0, 1)$ est un indicateur de sélection automatique de distribution. Il sépare principalement deux sous-classes de distributions semi-continues ($1 < p < 2$) et positives continues ($p \geq 2$). Nous explorons des outils de diagnostiques basés sur le test du rapport de vraisemblance et le test de Kolmogorov-Smirnov afin de discriminer des distributions très proches dans chaque sous-classe de ces deux modèles selon des valeurs de p. Basés sur l'unique égalité des indices de variation, nous discriminons également les distributions gamma et géométrique gamma avec $p = 2$ des familles Tweedie et géométriques Tweedie, respectivement. Nous effectuons une étude de simulation pour évaluer les procédures de discrimination dans ces sous-classes de deux familles. En se basant sur les probabilités de faire une sélection correcte, les distributions semi-continues ($1 < p \leq 2$) au sens large se distinguent nettement plus que les distributions continues sur-variées ($p > 2$). Pour terminer, deux jeux de données à des fins d'illustration sont étudiés.

Mots-clés. Distance Kolmogorov-Smirnov, Test du rapport de vraisemblance, Probabilité de faire une sélection correcte, Indice de variation, Indice de masse en zéro.

Abstract. In both Tweedie and geometric Tweedie models, the common power parameter $p \notin (0, 1)$ works as an automatic distribution selection. It mainly separates two subclasses of semicontinuous ($1 < p < 2$) and positive continuous ($p \geq 2$) distributions. We explore diagnostic tools based on the maximum likelihood ratio test and minimum Kolmogorov-Smirnov distance methods in order to discriminate very close distributions within each subclass of these two models according to values of p. Grounded on the unique equality of variation indices, we also discriminate the gamma and geometric gamma distributions with $p = 2$ in Tweedie and geometric Tweedie families, respectively. We perform a numerical comparison study to assess the discrimination procedures in these subclasses of two families. Based on probabilities of correct selection, semicontinuous ($1 < p \leq 2$) distributions in the broad sense are significantly more distinguishable than the over-varied continuous ($p > 2$) ones. Finally, two datasets for illustration purposes are investigated.

Keywords. Kolmogorov-Smirnov distance, Likelihood ratio test, Probability of correct selection, Variation index, Zero-mass index.

1 Introduction

Les modèles Tweedie et géométriques Tweedie fournissent des familles paramétriques flexibles de distributions pour traiter principalement des données asymétriques à droite et peuvent traiter des données continues avec une masse en zéro (Tweedie, 1984; Jørgensen et Kokonendji, 2011). Le paramètre de puissance commun $p \notin]0,1[$, appelé le paramètre de Tweedie est connecté à l'indice de stabilité (géométrique) commun $\alpha = (2-p)/(1-p)$, joue un rôle intrinsèque dans les deux modèles. En effet, p est un indicateur qui distingue chaque distribution dans chaque famille. La famille Tweedie comprend de nombreuses distributions spéciales, notamment gaussienne, Poisson, gamma décentrée, gamma et inverse gaussienne. La famille géométrique Tweedie, à son tour, provient de sommes géométriques de variables aléatoires Tweedie et peut être considérée comme un mélange exponentiel de la famille Tweedie (Abid et al., 2020). Des distributions particulières représentent la version géométrique de celles de Tweedie.

Comme préliminaires à une procédure de discrimination entre deux distributions, il est nécessaire que les deux distributions aient des caractéristiques communes telles que les supports et les allures des densités. Plus précisément, pour les familles de distributions de Tweedie et de géométriques Tweedie, nous considérerons les indices de masse en zéro et de variation récemment introduits par Abid et al. (2020) pour une variable aléatoire non négative Y. Rappelons que l'indice de masse en zéro est défini par $\mathrm{ZM}(Y) := \mathbb{P}(Y \leq y) \in [0,1]$ pour $y \to 0$. Ainsi, $\mathrm{ZM} \to \varrho$ lorsque $y \to 0$ désigne une distribution semi-continue si $\varrho > 0$ et une distribution absolument continue si $\varrho = 0$. Quant à l'indice de variation exprimé par $\mathrm{VI(Y)} = \mathrm{VarY}/(\mathbb{E}\mathrm{Y})^2 \in]0,+\infty[$, il est défini par rapport à la distribution exponentielle standard. En fait, Y est dite sur-, équi- et sous-variée par rapport à la loi exponentielle avec une moyenne $\mathbb{E}Y$ si $\mathrm{VI} > 1$, $\mathrm{VI} = 1$ et $\mathrm{VI} < 1$, respectivement.

L'idée de discriminer deux distributions a été initialement proposée dans le travail pionnier de Cox (1961). Et depuis, plusieurs auteurs ont abordé la discrimination entre deux distributions bien proches. La plupart de ces disriminations sont basés sur le test du rapport de vraisemblance maximale (LRT) et la distance minimale de Kolmogorov-Smirnov (KSD). L'objectif de cet article est de discriminer entre et dans les sous-classes des modèles Tweedie et géométriques Tweedie en utilisant les méthodes maximum LRT et minimum KSD. Ces modèles ont déjà été comparés dans le cadre des modèles linéaires généralisés (Kokonendji et al., 2020). Les sections 2 et 3 présentent certaines caractéristiques des deux modèles avec le cas commun de $p = 2$. La section 4 décrit les procédures proposées de discrimination et la probabilité estimée de faire une sélection correcte (PCS). La section 5 résume les résultats numériques et les axes d'application.

2 Propriétés de la famille Tweedie

Dans cette section, nous présentons certaines caractéristiques des modèles Tweedie continus et semi-continus. Soit X une variable aléatoire d'une distribution Tweedie, notée $Tw_p(m, \phi)$. Sa fonction de densité est donnée par

$$f_{Tw_p}(x; m, \phi) = a_p(x; \phi) \exp[\{x\psi_p(m) - K_p(\psi_p(m))\}/\phi] \mathbb{1}_{\mathbb{S}_p}(x), \tag{1}$$

où $\phi > 0$ est le paramètre de dispersion, $p \in]-\infty, 0] \cup [1, +\infty[$ est l'indice Tweedie déterminant la distribution, $\mathbb{S}_p$ est le support de la distribution, $a_p(x; \phi)$ est la fonction de normalisation, K_p est la fonction cumulative, ψ_p est la fonction inverse de K_p' et $m = K_p'(\theta)$ est la moyenne de X. Notons que $K_p'(\cdot)$ définit un difféomorphisme entre son domaine canonique Θ_p et son image $M_p := K_p'(\Theta_p)$ qui est son domaine des moyennes. Bien que les densités de Tweedie ne sont généralement pas explicites, leurs fonctions cumulantes sont simples. Les deux ensembles $\mathbb{S}_p$ et M_p dépendent de p. Pour $p = 0, p = 1$, $1 < p < 2$ et $p \geq 2$, le support consiste à la droite réelle $\mathbb{R}$, des entiers naturels $\mathbb{N}$, des réelles positives ou nulles $[0, +\infty[$ et strictement positives $]0, +\infty[$, respectivement. Les domaines des moyennes dans ces cas sont les supports convexes de $\mathbb{S}_p$ correspondants. Néanmoins, pour $p < 0$, on a $\mathbb{S}_p = \mathbb{R}$ et $M_p =]0, +\infty[$. Tableau 1 présente les sous-classes des modèles Tweedie.

Modèles (géométriques) Tweedie	$\alpha = \alpha(p)$	p	$\mathbb{S}_p$	M_p
(Géométrique) Extrême stable	$1 < \alpha < 2$	$p < 0$	$\mathbb{R}$	$]0, +\infty[$
(Laplace asymétrique/) Gaussien	$\alpha = 2$	$p = 0$	$\mathbb{R}$	$\mathbb{R}$
[N'existe pas]	$\alpha > 2$	$0 < p < 1$		
(Géométrique) Poisson	$\alpha = -\infty$	$p = 1$	$\mathbb{N}$	$]0, +\infty[$
(Géométrique) Poisson-gamma-composé	$\alpha < 0$	$1 < p < 2$	$[0, +\infty[$	$]0, +\infty[$
(*Géométrique*) *gamma décentré*	$\alpha = -1$	$p = 3/2$	$[0, +\infty[$	$]0, +\infty[$
`(Géométrique) Gamma`	$\alpha = 0$	$p = 2$	$]0, +\infty[$	$]0, +\infty[$
(Géométrique Mittag-Leffler/) Positive stable	$0 < \alpha < 1$	$p > 2$	$]0, +\infty[$	$]0, +\infty[$
(*Ressel-Kendall/*) *Inverse Gaussien*	$\alpha = 1/2$	$p = 3$	$]0, +\infty[$	$]0, +\infty[$

Tableau 1: Résumé des modèles Tweedie et de géométriques Tweedie, y compris leur indice de stabilité commun $\alpha = \alpha(p)$, puissance p, support des distributions $\mathbb{S}_p$ et domaine des moyennes M_p.

Étant donnée la moyenne m de $X \sim Tw_p(m, \phi)$, sa variance est ϕm^p. Ainsi,

$$\mathrm{VI}(Tw_p) = \phi m^{p-2} \quad \left(\gtreqless 1 \iff \phi \gtreqless m^{2-p} \right). \tag{2}$$

Les comportements de $\mathrm{VI}(Tw)$ dans (2) sont des sur-variations pour tous $p \notin]0, 1]$ et une équi-variation pour $p = 2$. Le cas spécial de $\mathrm{VI}(Y) = \phi$ dans (2) correspondant à la distribution gamma ($p = 2$) ne dépend pas de la moyenne m.

3 Propriétés de la famille géométrique Tweedie

Pour cette section, nous nous intéressons aux modèles géométriques Tweedie continus et semi-continus résultant des sommes géométriques des variables de Tweedie. Soit $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ la variable géométrique Tweedie de paramètre de puissance $p \notin]0,1[$, de paramètre de dispersion $\widetilde{\phi} > 0$ et de moyenne $\widetilde{m}$. On a donc la représentation suivante :

$$Z = \sum_{j=1}^{G} T_j,$$

où $T_1, T_2, \ldots$ sont des variables Tweedie indépendantes et identiquement distribuées à $Tw_p(m, \phi)$ et G est une variable aléatoire géométrique, indépendante des T_j, avec la fonction de masse de probabilité $\mathbb{P}(G = g) = q(1-q)^{g-1}$, pour $g = 1, 2, \ldots$ et $q \in]0,1[$. De plus, la famille géométrique Tweedie est intérprétée comme un mélange exponentiel (voir, par exemple, Abid et al., 2020, Proposition 2.1) et elle est donc exprimée par la formulation hiérarchique suivante

$$X \sim \text{Exponentielle}(1) \quad \text{et} \quad Z|(X = x) \quad \sim Tw_p(x\widetilde{m}, x^{1-p}\widetilde{\phi}).$$

La fonction de densité de $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ est donnée en fonction de (1) par

$$f_{GTw_p}(z; \widetilde{m}, \widetilde{\phi}, p) = \int_0^\infty \exp(-x) f_{Tw_p}(z; x\widetilde{m}, x^{1-p}\widetilde{\phi}) dx.$$

Cette densité n'a toujours pas de forme explicite, sauf pour $p \in \{0, 1, 2, 3\}$. La méthode de Monte Carlo fournit une approximation très raisonnable $\widehat{f_{GTw_p}}$ de f_{GTw_p}, grâce à la disponibilité de la densité de Tweedie f_{Tw_p} via la fonction R `dtweedie`.

Etant donné la moyenne $\widetilde{m}$ de $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$, sa variance est $\widetilde{m}^2 + \widetilde{\phi}\widetilde{m}^p$. D'où,

$$\text{VI}(GTw_p) = 1 + \widetilde{\phi}\widetilde{m}^{p-2} \quad \left(\gtreqless 1 \;\Leftrightarrow\; \widetilde{\phi} \gtreqless 0 \right). \tag{3}$$

En considérant la possibilité d'obtenir $\widetilde{\phi}$, les comportements de VI(GTw) dans (3) des modèles géométriques Tweedie étendus sont clairement sur-, équi- et sous-variations pour $p \notin]0,1]$, $p = 2$ et $p \in]-\infty, 0] \cup]1, 2]$, respectivement. Toutefois, la fonction densité associée f_{GTw_p} n'existe pas pour $\widetilde{\phi} < 0$. Notons que, tout comme les modèles Tweedie avec $p = 2$ dans (2), l'indice de variation VI(GTw) dans (3) pour le cas particulier $p = 2$ correspondant à la distribution géométrique gamma est égal à $1 + \widetilde{\phi}$ et ne dépend pas de la moyenne $\widetilde{m}$. Pour $p = 2$ et étant donnés les modèles $\widetilde{m} = m > 0$, les deux indices de variation pour Tweedie (2) et géométrique Tweedie (3) coïncident lorsque leurs paramètres de dispersion diffèrent de +1 au sens de géométrique Tweedie. Plus conventionnellement, on peut écrire $Tw_2(m, \phi) \approx GTw_2(m, 1 + \phi)$ pour $\phi \geq 1$ et tout $m > 0$.

4 Procédures de discrimination

Pour diagnostiquer le modèle d'ajustement approprié parmi deux distributions données pour un jeu de données, deux techniques sont envisagées impliquant le maximum LRT et le minimum KSD comme critères d'optimalité. Considérons un échantillon aléatoire $Y_1, Y_2, \ldots, Y_n$ qui appartient à l'une des distributions parentes $f_p(y; m, \phi)$. Pour $p > 1$ fixé, les estimateurs du maximum de vraisemblance de la moyenne et du paramètre de dispersion sont donnés par

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{et} \quad \widehat{\phi} = \arg\max_{\phi>0} L_p(\widehat{m}, \phi),$$

où $L_p(\widehat{m}, \phi)$ est la fonction de vraisemblance profilée calculée en $\widehat{m}$. La statistique du rapport de vraisemblance, également appelée la statistique de Cox, est définie par

$$LT_{p_j,p_{j'}} = \log\left(\frac{L_{p_j}(\widehat{m}_j, \widehat{\phi}_j)}{L_{p_{j'}}(\widehat{m}_{j'}, \widehat{\phi}_{j'})}\right).$$

La règle de décision pour discriminer entre deux distributions ayant des densités f_{p_j} et $f_{p_{j'}}$ est de choisir f_{p_j} si $LT_{p_j,p_{j'}} > 0$, et de rejeter f_{p_j} en faveur de $f_{p_{j'}}$ sinon. Notons que, contrairement au LRT, le test KSD peut considérer plus de deux distributions compétitives pour décrire les données. Le KSD est, quant à lui, défini par

$$KS_{p_j} = \sup_{-\infty<y<+\infty} |\widehat{F}_{p_j}(y; \widehat{m}_j, \widehat{\phi}_j) - \widetilde{F}(y)|, \quad j \in \{1, \ldots, \ell\},$$

avec $\ell \geq 2$, $\widehat{F}_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ la fonction de distribution de $f_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ et $\widetilde{F}(\cdot)$ la fonction de distribution empirique calculée directement à partir des données. L'indice du modèle j_0 ayant la distance minimale est donc sélectionné comme modèle gagnant :

$$j_0 = \arg\min_{j\in\{1,\ldots,\ell\}} KS_{p_j}.$$

Les performances des méthodes maximum LRT et minimum KSD sont étudiées par les PCS à partir de simulations. En pratique, nous générons $(Y_n^{(1)}, \ldots, Y_n^{(N)})$, où $Y_n^{(k)}$ sont k-échantillons aléatoires de taille n qui appartiennent à f_p. Nous répétons les deux procédures, LRT et KSD, pour chaque $Y_n^{(k)}$, $k = 1, \ldots, N$. Le PCS qui correspond à la proportion de fois f_p est choisi comme modèle gagnant et peut être évalué par :

$$\widehat{PCS}_p = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}\{Y_n^{(k)} \text{ est correctement classifié}\}.$$

5 Simulations et applications

Nous appliquerons les méthodes LRT et KSD pour discriminer entre les modèles communs de Tweedie et géométriques Tweedie d'une part et dans chaque sous-classes des modèles Tweedie et géométriques Tweedie d'autre part. D'abord, nous considérons la discrimination entre gamma $Tw_2(m, \phi)$ et géométrique gamma $GTw_2(\widetilde{m}, \widetilde{\phi})$ vérifiant $\widetilde{\phi} = 1 + \phi$. Puis, nous supposons que la distribution parente est Tw_p et les distributions alternatives sont $Tw_{p+\varepsilon}$, avec $\varepsilon > 0$ tel que Tw_p et $Tw_{p+\varepsilon}$ sont de même type (voir Table 1). Ce qui vise à détecter l'évolution de la discrimination entre les distributions pour chaque type : $1 < p < 2$ and $p > 2$. Finalement, nous supposons que la distribution parente est GTw_p et les distributions alternatives sont $GTw_{p+\varepsilon}$, avec $\varepsilon > 0$ tel que GTw_p et $GTw_{p+\varepsilon}$ sont de même type.

Nous comparons dans chaque configuration l'évolution du PCS pour différentes combinaisons de paramètres et de tailles d'échantillons. La méthode LRT s'est avérée plus performante que KSD. De plus, les distributions semi-continues ($1 < p \leq 2$) au sens large sont nettement plus distinguables que celles continues sur-variées ($p > 2$) des deux familles respectives. Nous analysons deux jeux de données. Concernant le premier, les distributions gamma Tw_2 et géométrique gamma GTw_2 sont comparées. Pour le second, les deux sous-classes semi-continues ($1 < p < 2$) de Tweedie et géométrique Tweedie sont considérées en suggérant différentes valeurs de p.

Bibliographie

Abid, R., Kokonendji, C. C. and Masmoudi, A. (2020). Geometric-Tweedie regression models for continuous and semicontinuous data with variation phenomenon, *AStA Advances in Statistical Analysis*, 104, 33-58.

Cox, D. R. (1961). Tests of separate families of hypotheses, Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 105-123.

Jørgensen, B. and Kokonendji, C. C. (2011). Dispersion models for geometric sums. *Brazilian Journal of Probability and Statistics*, 25, 263-293.

Kokonendji, C. C., Bonat, W. H. and Abid, R. (2020). Tweedie regression models and its geometric sums for (semi-)continuous data. *WIREs Computational Statistics*, 12, in press (DOI : 10.1002/WICS.1496).

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics : Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* (J. K. Ghosh and J. Roy, eds.), pp. 579-604, Indian Statistical Institute, Calcutta.

Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering

Julien Ah-Pine [1,2]

[1] *Université de Lyon, Lyon 2 et ERIC EA3083, 5 Avenue Pierre Mendèes France, 69500 Bron, France; julien.ah-pine@univ-lyon2.fr*
[2] *Université Clermont Auvergne, LMBP UMR6620, 3 place Vasarely, 63170 Aubière, France*

Résumé. Nous nous intéressons à la tâche de clustering du point de vue graphe à l'instar du partionnement spectral (spectral clustering). Dans ce cas, la matrice d'affinité qui mesure l'intensité du lien (arête du graphe) pour chaque paire d'éléments (sommets du graphe) joue un rôle crucial. Plusieurs travaux antécédents ont montré l'intérêt de transformer une matrice d'affinité initiale de sorte à satisfaire certaines propriétés. La bistochasticité est une condition pertinente à cet égard. Dans ce travail, nous mettons en avant une autre condition: l'idempotence. Par la suite, En utilisant les propriétés existantes entre les matrices bistochastiques et idempotentes d'une part, et leurs matrices Laplaciennes associées d'autre part, nous proposons une nouvelle méthode d'apprentissage non-supervisé de matrice d'affinité. Notre procédure d'optimisation repose sur la méthode des multiplicateurs de Lagrange avec directions alternées (ADMM). Des résultats expérimentaux montrent l'intérêt pratique de notre approche.

Mots-clés. Clustering, Matrice d'affinité, Bistochasticité, Idempotence, ADMM.

Abstract. We are interested in graph based clustering such as spectral clustering. In this context, the affinity matrix that provides the strenght of the similarity between each pair of elements plays a crucial role. Several previous works have showed that transforming a given affinity matrix so that it becomes double stochastic was beneficial. In this work, we highlight another property: idempotency. By leveraging the relationships between double stochastic and idempotent matrices on the one hand, and their related Laplacian matrices on the other hand, we introduce a new unsupervised learning method for affinity matrices. Our learning algorithm is based on ADMM. Some experimental results are provided in order to demonstrate the interest of our proposal.

Keywords. Clustering, Affinity matrix, Double stochasticity, Idempotency, ADMM.

1 Contexte et travaux antérieurs

La tâche de clustering consiste à partitionner un ensemble d'éléments en des sous-ensembles homogènes appelés clusters. Soit un ensemble de n vecteurs $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ appartenant à

$\mathbb{R}^p$, que nous cherchons à analyser. Nous nous intéressons à la partition en k clusters $C = \{C_1, \ldots, C_k\}$ qui minimise le critère SSE (Sum of Squared Errors) suivant :

$$\sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\phi(\mathbf{x}_i) - \mathbf{c}_j\|^2 \tag{1}$$

où $\phi : \mathbb{R}^p \to \mathbb{F}$ est une projection des $\{\mathbf{x}_i\}$ dans un espace de grande dimension $\mathbb{F}$, $\mathbf{c}_j = \sum_{\mathbf{x}_i \in C_j} \phi(\mathbf{x}_i)/n_j$ est le vecteur moyen du cluster C_j qui est de cardinal n_j et $\|.\|$ est la norme Euclidienne dans $\mathbb{F}$.

Le critère SSE peut être formalisé à l'aide de la matrice de noyau $\mathbf{K}$ de terme général $\mathbf{K}_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle = \kappa(\mathbf{x}_i, \mathbf{x}_{i'})$ où $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ est une fonction noyau. Il s'agit dans ce cas de déterminer $\mathbf{X}$, la matrice de $\mathbb{R}^{n \times n}$ qui minimise la fonction objectif suivante:

$$\mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{2}$$

où Tr est l'application trace dans $\mathbb{R}^{n \times n}$, $\mathbf{I}_n$ est la matrice identité d'ordre n, et $\mathbf{X}$ est de terme général:

$$\mathbf{X}_{ii'} = \begin{cases} 1/n_j & \text{si } \mathbf{x}_i \text{ et } \mathbf{x}_{i'} \text{ sont dans } C_j, \\ 0 & \text{sinon.} \end{cases} \tag{3}$$

La matrice $\mathbf{X}$ ainsi définie possède plusieurs propriétés. Plus précisément [5] montre que la minimisation du SSE peut s'exprimer de façon équivalente comme suit:

$$\begin{aligned} &\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \\ &\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2, \mathrm{Tr}(\mathbf{X}) = k. \end{aligned} \tag{4}$$

où $\mathbf{0}_n$ est la matrice nulle d'ordre n, $(\mathbf{X} \geq \mathbf{0}_n) \Leftrightarrow (\mathbf{X}_{ii'} \geq 0, \forall i, i' = 1, \ldots, n)$, $\mathbf{X}^\top$ est la matrice transposée de $\mathbf{X}$ et $\mathbf{e}_n$ est le vecteur rempli de 1 de dimension n.

La matrice $\mathbf{X}$ recherchée est ainsi non-négative, symétrique, bistochastique (les sommes de chaque ligne et de chaque colonne valent 1), idempotente et de trace égale à k le nombre de clusters désiré. En fait, il existe une bijection entre l'ensemble des partitions d'un ensemble de n éléments et l'ensemble des classes d'équivalence des matrices bistochastiques et idempotentes pour la relation $\mathbf{X} \sim \mathbf{Y}$ si et seulement si il existe une matrice de permutation $\mathbf{P}$ telle que $\mathbf{X} = \mathbf{P}\mathbf{Y}\mathbf{P}^\top$ [6].

Par exemple la partition $\{(a, e), (b, c, d)\}$ peut être représentée par les matrices bistochastiques et idempotentes suivantes appartenant à la même classe d'équivalence:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{pmatrix} \end{array} \sim \begin{array}{c} \\ a \\ e \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} a & e & b & c & d \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

Le problème ainsi formulé, donne un point de vue graphe à la tâche de clustering: $\mathbf{K}$ (à condition d'être non-négative -comme pour le noyau Gaussien par exemple-) peut être vue telle une matrice d'adjacence pondérée d'un graphe sans structure particulière et $\mathbf{X}$ peut être interprétée comme la matrice d'adjacence pondérée d'un graphe représentant une partition des sommets. Il s'agit alors d'approximer $\mathbf{K}$ par $\mathbf{X}$ au sens de la norme de Frobenius. En effet, soit $\mathrm{Tr}(\mathbf{X}^\top \mathbf{Y}) = \langle \mathbf{X}, \mathbf{Y} \rangle_F$ le produit scalaire de Frobenius dans $\mathbb{R}^{n \times n}$. Si $\mathbf{X}$ vérifie les contraintes stipulées dans (4), alors il est facile de montrer que:

$$\min_{\mathbf{X}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \Leftrightarrow \min_{\mathbf{X}} \|\mathbf{K} - \mathbf{X}\|_F^2 \tag{5}$$

Le modèle d'optimisation (4), appelé 0-1 Semi-Definite Program dans [5], est NP-difficile en raison de la nature discrète et de l'idempotence de la matrice $\mathbf{X}$. En pratique, une démarche heuristique permettant de résoudre de façon approchée (4) consiste à (i) définir un problème relaxé solutionnable en temps polynomial, (ii) discrétiser la solution optimale du problème relaxé afin d'obtenir une solution réalisable du problème initial.

De nombreuses méthodes d'apprentissage de matrice d'affinité et de clustering découlent de cette démarche. Dans ce travail nous nous penchons plus particulièrement, sur les approches présentées dans [9] et [10]. Ces travaux reviennent à remplacer $\mathbf{K}$ par une matrice $\mathbf{X}$ non-négative, symétrique et bistochastique (étape (i)). Autrement dit, la contrainte d'idempotence est abandonnée, le nombre de cluster k n'est alors plus associé à la trace et les contraintes restantes sont toutes linéaires. Des procédures efficaces sont proposées pour déterminer $\mathbf{X}$: dans [9] il s'agit d'une version symétrique de l'algorithme de Sinkhorn-Knoop [6] dénoté SSK, alors que dans [10] est introduit l'algorithme DSN (Double Stochastic Normalization). Plus précisément, l'algorithme de Sinkhorn-Knoop vise à minimiser la divergence de Kullback-Leibler entre $\mathbf{K}$ et $\mathbf{X}$ alors que l'approche DSN résulte de la minimisation de la distance de Frobenius entre $\mathbf{K}$ et $\mathbf{X}$. Une fois $\mathbf{X}$ déterminée une méthode de discrétisation est utilisée. Le spectral clustering qui, en bref, applique l'algorithme des k-means sur les k premiers vecteurs propres de $\mathbf{X}$ est une approche classique à cet égard (étape (ii)).

2 Approche proposée

Contrairement aux deux approches précédentes, nous cherchons à tenir compte de la contrainte d'idempotence tout en évitant de se ramener à un problème NP-difficile. Néanmoins, nous ne considérons pas le nombre de clusters k comme paramètre de notre modèle et n'imposons donc pas $\mathrm{Tr}(\mathbf{X}) = k$. L'approximation basée sur la distance de Frobenius reste centrale dans notre approche qui vise, en somme, à définir un problème relaxé du

modèle suivant:

$$\begin{aligned} &\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{K}-\mathbf{X}\|_F^2 \\ &\text{s.t. } \mathbf{X}\geq \mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n, \mathbf{X}=\mathbf{X}^2. \end{aligned} \tag{6}$$

$\mathbf{X}$ étant bistochastique, la matrice des degrés vaut $\mathbf{I}_n$ et la matrice Laplacienne associée à $\mathbf{X}$ est donnée par $\mathbf{L}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}$. Clairement, le problème (6) peut être reformulé de façon équivalente en fonction de $\mathbf{L}_\mathbf{X}$ comme suit :

$$\begin{aligned} &\min_{\mathbf{L}_\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{I}_n-\mathbf{K}-\mathbf{L}_\mathbf{X}\|_F^2 \\ &\text{s.t. } \mathbf{L}_\mathbf{X}\leq \mathbf{I}_n, \mathbf{L}_\mathbf{X}=\mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n=\mathbf{n}_n, \mathbf{L}_\mathbf{X}=\mathbf{L}_\mathbf{X}^2. \end{aligned} \tag{7}$$

où $\mathbf{n}_n$ est le vecteur nul de dimension n.

Nous exploitons à présent les relations algébriques existantes entre $\mathbf{X}$ et $\mathbf{L}_\mathbf{X}$. En effet, ces deux matrices étant symétriques et idempotentes, elles représentent des projections orthogonales. De façon plus singulière, l'une est l'unique projecteur orthogonale complémentaire de l'autre et *vice-versa*: l'image de $\mathbf{X}$ est le noyau de $\mathbf{L}_\mathbf{X}$, l'image de $\mathbf{L}_\mathbf{X}$ est le noyau de $\mathbf{X}$ et nous avons la relation suivante, centrale dans notre travail:

$$\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}_n \tag{8}$$

Nous proposons de relaxer (6) et (7) en considérant le modèle suivant:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{L}_\mathbf{X}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{K}-\mathbf{X}\|_F^2 + \frac{1}{2}\|\mathbf{I}_n-\mathbf{K}-\mathbf{L}_\mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2 \\ &\text{s.t. } \begin{cases} \mathbf{X}\geq \mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n, \\ \mathbf{L}_\mathbf{X}\leq \mathbf{I}_n, \mathbf{L}_\mathbf{X}=\mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n=\mathbf{n}_n, \\ \mathbf{X}+\mathbf{L}_\mathbf{X}=\mathbf{I}_n. \end{cases} \end{aligned} \tag{9}$$

où $\mu > 0$ est un paramètre de pénalité.

Notre modèle intitulé DSNI (Doubly Stochastic and Nearly Idempotent), consiste en un apprentissage joint de $\mathbf{X}$ et de sa matrice Laplacienne associée $\mathbf{L}_\mathbf{X}$. Il est facile de montrer que sous la condition $\mathbf{X}+\mathbf{L}_\mathbf{X} = \mathbf{I}_n$, les trois propriétés qui suivent sont équivalentes: $\mathbf{X}=\mathbf{X}^2$, $\mathbf{L}_\mathbf{X}=\mathbf{L}_\mathbf{X}^2$, $\mathbf{X}\mathbf{L}_\mathbf{X}=\mathbf{0}_n$. Cependant, ces propriétés étant la source première de la complexité des problèmes (6) et (7), nous ne les intégrons pas dans les contraintes de notre modèle. Pour pallier à ce manque, nous ajoutons, en revanche, un terme de pénalité dans la fonction objectif, $\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2$, afin d'encourager les solutions obtenues à être ainsi quasi-idempotentes.

Le problème DSNI (9) étant bi-convexe, nous pouvons utiliser la méthode ADMM (voir par exemple [2]) comme procédure d'optimisation. Les différentes étapes sont alors les suivantes:

0. Initialisation: $\mathbf{X}^0 \leftarrow \mathbf{K}$ (en ayant au préalable annuler les valeurs négatives de $\mathbf{K}$ le cas échéant).

1. Déterminer $\mathbf{L}_{\mathbf{X}}^{t+1}$ avec $\mathbf{X}^t$ fixé:

$$\mathbf{L}_{\mathbf{X}}^{t+1} \leftarrow \underset{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}^t\mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\rho}{2}\|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2$$
$$\text{s.t. } \mathbf{L}_{\mathbf{X}} \leq \mathbf{I}_n, \mathbf{L}_{\mathbf{X}} = \mathbf{L}_{\mathbf{X}}^\top, \mathbf{L}_{\mathbf{X}}\mathbf{e}_n = \mathbf{n}_n. \tag{10}$$

2. Déterminer $\mathbf{X}^{t+1}$ avec $\mathbf{L}_{\mathbf{X}}^{t+1}$ fixé:

$$\mathbf{X}^{t+1} \leftarrow \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 + \frac{\rho}{2}\|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2$$
$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n. \tag{11}$$

3. Déterminer $\mathbf{U}^{t+1}$:

$$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{X}^{t+1} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n \tag{12}$$

4. Répéter 1., 2., 3. tant qu'une condition d'arrêt n'est pas satisfaite.

Les sous-problèmes (10) et (11) peuvent être résolus efficacement par projections successives sur des ensembles convexes (voir par exemple [1]).

3 Validation empirique de l'approche proposée

Nous avons testé notre approche sur plusieurs jeux de données réels classiques disponibles en ligne[1]. Le protocole expérimental est le suivant: calcul de $\mathbf{K}$ en utilisant un noyau Gaussien; approximation de $\mathbf{K}$ par une matrice d'affinité bistochastique $\mathbf{X}$ obtenue par SSK ou DSN ou DSNI (sauf pour la baseline); application du spectral clustering [3] sur $\mathbf{X}$ en fixant k au nombre correct de clusters; comparaison de la partition obtenue et de la vérité terrain en utilisant la mesure NMI (Normalized Mutual Information). Pour le noyau Gaussien, l'hyperparamètre σ^2 est fixé à p et pour DSNI le paramètre de pénalité μ est fixé à $\sqrt{n}$.

Les résultats obtenus sont donnés dans la Table 1. La colonne SC représente la baseline et utilise le spectral clustering directement sur $\mathbf{K}$ la matrice de noyau Gaussien. Sur l'ensemble des jeux de données, nous constatons que DSNI donne de meilleurs résultats que SSK et DSN ce qui valide l'intérêt de notre modèle.

[1] `https://archive.ics.uci.edu/ml/index.php`

Dataset	n	p	k	SC	SSK	DSN	DSNI
Glass	214	9	6	0.253	0.276	0.243	**0.297**
Ionosphere	351	34	2	0.038	0.066	0.076	**0.131**
Breast cancer	569	30	2	0.010	0.010	0.010	**0.670**
Yeast	1484	8	10	0.070	0.258	0.256	**0.263**
Digits	1797	64	10	0.015	0.044	0.743	**0.767**

Table 1: Statistiques des jeux de données et mesures NMI des différentes méthodes.

References

[1] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

[2] S. Boyd, N. Parikh, and E. Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[3] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[4] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1):186–205, 2007.

[5] J. Peng and Y. Xia. A new theoretical framework for k-means-type clustering. In *Foundations and advances in data mining*, pages 79–96. Springer, 2005.

[6] R. Sinkhorn. Two results concerning doubly stochastic matrices. *The American Mathematical Monthly*, 75(6):632–634, 1968.

[7] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[8] F. Wang, P. Li, and A. C. Konig. Learning a bi-stochastic data similarity matrix. In *2010 IEEE International Conference on Data Mining*, pages 551–560. IEEE, 2010.

[9] R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 294–301. IEEE, 2005.

[10] R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in neural information processing systems*, pages 1569–1576, 2007.

Méthode de comparaison d'aires sous la courbe dans des essais cliniques avec arrêt prématuré du suivi : application aux vaccins thérapeutiques contre le VIH

Marie Alexandre[1] & Mélanie Prague[2] & Rodolphe Thiébaut[3]

Inria SISTM Team, INSERM U1219, Université de Bordeaux, ISPED - France
Vaccine Research Institute, Créteil - France
[1] *marie.alexandre@inria.fr,* [2] *melanie.prague@inria.fr,* [3] *rodolphe.thiebaut@inria.fr*

Résumé. Les interruptions de traitement analytiques (ATI) sont couramment utilisées pour évaluer l'efficacité de nouveaux vaccins thérapeutiques contre le VIH. Ces procédures nécessitent alors la détermination de critères de jugement synthétiques permettant de rendre compte de cette efficacité par comparaison entre différents bras de vaccination tels que l'aire sous la courbe de charge virale moyennée sur le temps de suivi (nAUC). Cependant, dû à la nécessité de remettre les patients à risque sous traitement, l'existence de données manquantes au hasard (MAR) monotone est inévitable au cours d'ATI pouvant mener à des résultats biaisés des tests de comparaison. Cette étude a pour objectif de présenter une méthode évaluant la différence de nAUC entre deux bras de vaccination à partir des dynamiques marginales de groupes estimées par des modèles à effets mixtes. L'application de cette méthode sur des simulations d'essais randomisés à deux bras de vaccination a été menée afin d'en vérifier les propriétés statistiques et de montrer sa supériorité vis-à-vis de méthodes adhoc communément utilisées ainsi que d'un test non-paramétrique. Son application sur données réelles a permis de confirmer cette conclusion.
Mots-clés. VIH, efficacité vaccinale, AUC, données manquantes, modèle à effets mixtes, test statistiques

Abstract. Analytic treatment interruption (ATI) are commonly used to evaluate new HIV therapeutic vaccins efficacy. These protocols require the choice of summary endpoints, such as the area under the HIV RNA load curve normalized by follow-up time (nAUC), to assess this efficacy by comparing them between different vaccine arms. However, monotonic missing at random (MAR) data are unavoidable during ATI leading to potential biased results for comparison tests. This study aimed to present a method evaluating the difference of nAUC between two vaccine arms based on marginal dynamics of group level estimated by mixed effects models. In order to evaluate its statistical properties, the method was applied on simulated two-armed randomized vaccine trials where the difference of AUC between the two vaccine arms as well as the missingness were varied. Simuations allowed to show its supriority compared to commonly used adhoc approaches and non-parametric test. Its application on real data allowed to confirm this conclusion.
Keywords. HIV, vaccine efficacy, AUC, missing data, mixed effects models, statistical test

1 Introduction

1.1 Contexte

Le développement de vaccins thérapeutiques est un aspect important dans la recherche de stratégies du contrôle viral du VIH à long terme. Ces derniers ont pour objectif la diminution voir l'élimination complète de l'infection virale jusqu'à présent rendu impossible par l'existence d'un réservoir viral persistant. L'absence de biomarqueurs reconnus capable de prédire le contrôle virologique en l'absence de traitement antiretroviraux (ART) fait des interruptions de traitement analytiques (ATI) l'unique moyen d'évaluer la capacité d'une nouvelle stratégie à contrôler la virémie après arrêt d'ART. Dans ce type d'étude, un choix judicieux de critère de jugement virologique est l'aire sous la courbe de charge virale normalisée par le temps de suivi (nAUC).

1.2 Impact des données manquantes sur la statistique de test

Ces essais sont régis par des critères éthiques stricts afin de minimiser les risques encourus par les patients, tels que des critères de reprise prématurée des ART. Ces reprises précoces de traitement au cours de l'ATI, basées sur des règles définies dans le protocole, notamment pour éviter des niveaux de charges virales trop élevées pour garantir un risque minimal aux patients, sont traitées comme une sortie de l'étude et génère par conséquent des données manquantes monotones. Notre connaissance quasi déterministique du processus d'exclusion, tel que le niveau de charge virale maximal autorisé, attribue le caractère manquantes au hasard (MAR) à ces données. D'un point de vue statistique, la non-prise en compte des données manquantes dans des tests classiques d'égalité de moyennes de nAUCs mène à de mauvaises propriétés statistiques telles que des erreurs de type-I élevées, des pertes de puissance ou encore un biais sur les résultats du test [1].

2 Objectifs

Nous avons pour objectif de proposer une méthodologie statistique pour tester l'efficacité de ces vaccins en comparant les dynamiques de charge virales entre les différents bras de vaccination. A ces fins, nous nous basons sur un critère de jugement facilement mesurable, précis et interprétable capable de résumer les dynamiques de charge virale, l'AUC normalisée sur la période d'interruption de traitement (nAUC). En 2014, Bell. et al [2] ont mis en évidence l'intérêt d'utiliser des méthodes basées sur le maximum de vraisemblance pour réduire le biais induit par les données manquantes de type MAR et MNAR dans le calcul de l'AUC. En se basant sur ces résultats, nous construisons un test statistique, construit sur les dynamiques marginales de nos données longitudinales estimées par un modèle à effets mixtes, permettant de conclure de l'existence d'une différence de nAUC entre nos deux groupes d'intérêt en présence de données MAR monotones.

3 Méthodes

3.1 Définition du nAUC par méthode d'interpolation

On considère un essai clinique comptabilisant N patients répartis au sein de G groupes de vaccination. On note $Y_{ij,g}$ la mesure de charge virale du sujet i appartenant au group g à son $j$ème temps de mesure, $i \in \{1, \cdots, N\}$, $g \in \{1, \cdots, G\}$. En définissant $\{t_{j,g}\} = \cup_{i \in g}(\{t_{ij,g}\})$ comme l'ensemble des temps de mesures observés dans le groupe g avec $j \in \{1, \cdots, m_g\}$, on définit le nAUC à l'échelle des groupes, et son approximation par la méthode des trapèzes communément utilisée, par

$$n\text{AUC}_g = \frac{1}{T_g}\int_0^{T_g} \overline{Y}_g(t)dt \simeq \frac{1}{T_g}\sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2}(\overline{Y}_{j,g} + \overline{Y}_{j-1,g})$$

où $\overline{Y}_g$ correspond à la moyenne des observations dans le groupe et T_g au temps de suivi.

3.2 Estimation du nAUC par modèle à effets mixtes

On considère un modèle à effets mixtes linéaire (MEM) pour décrire nos données d'intérêt $\boldsymbol{Y}$ définie par la formulation matricielle $\boldsymbol{Y} = \boldsymbol{X_0\gamma} + \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$ où $\boldsymbol{X}$ et $\boldsymbol{X_0}$ sont respectivement les matrices de design pour les effets fixes spécifiques et non-spécifiques aux groupes et $\mathbf{Z}$ celles des effets aléatoires avec $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ et $\boldsymbol{b}$ leurs vecteurs de coefficients de régression respectifs. On suppose le modèle d'erreur ainsi que les effets aléatoires indépendants, centrés, normalement distribués de variances respectives $\boldsymbol{\Theta}$ et $\boldsymbol{\Omega}$. Par construction, la matrice $\boldsymbol{X}$ et le vecteur $\boldsymbol{\beta}$ sont définis comme $\boldsymbol{X} = \text{diag}(\boldsymbol{X_1}, \cdots, \boldsymbol{X_G})$ et $\boldsymbol{\beta^T} = (\boldsymbol{\beta^{1^T}}, \cdots, \boldsymbol{\beta^{G^T}})$ avec $\boldsymbol{X_g}$ et $\boldsymbol{\beta^g}$, les grandeurs spécifiques à chaque groupe. En définissant l'estimation de la dynamique marginale du groupe g par $\widehat{\boldsymbol{\mu}}_{\boldsymbol{g}} = \mathbb{E}(\widehat{\boldsymbol{Y}}_{\boldsymbol{g}}) = \boldsymbol{X_0^{[g]}}\widehat{\boldsymbol{\gamma}} + \boldsymbol{X_g}\widehat{\boldsymbol{\beta}}^{\boldsymbol{g}}$ et les poids de l'approximation de l'intégrale par $\boldsymbol{\omega_g} = (w_{1,g}, \cdots, w_{m_g,g})^T$, l'approximation du nAUC du groupe g peut alors s'exprimer comme

$$\widehat{n\text{AUC}}_g = \frac{1}{T_g}\boldsymbol{w_g^T}\widehat{\boldsymbol{\mu}}_g \qquad\qquad w_{j,g} = \begin{cases} \frac{t_{j+1,g} - t_{j,g}}{2}, & j = 1 \\ \frac{t_{j,g} - t_{j-1,g}}{2}, & j = m_g \\ \frac{t_{j+1,g} - t_{j-1,g}}{2}, & \text{sinon} \end{cases}$$

3.3 La statistique de test

On veut identifier si deux groupes de traitement distincts peuvent être différencier par leur valeur moyenne d'aire sous la courbe. Par conséquent, on définit nos hypothèses d'intérêt pour deux groupes comparés g_1 et g_2 comme l'égalité (H_0) et la différence (H_1) de leur nAUC. La présence de censure informative par sortie d'étude impactant le temps

de suivi au sein et entre chaque groupe, la statistique de test est construite de manière à comparer les nAUC de g_1 et g_2 sur le même intervalle de temps. Pour cela, on note $T = \min(T_{g_1}, T_{g2})$ et on définit $\widehat{n\text{AUC}}_g^{\text{rest}} = \widehat{n\text{AUC}}_g\Big|_{T_g=T}$ pour $g \in \{g_1, g_2\}$. De part les hypothèses de normalité des modèles à effets mixtes et leur estimation par des méthodes de maximum de vraisemblance, on peut construire la Z-statistique normalement distribuée donnée par

$$Z = \frac{\widehat{n\text{AUC}}_{g_2}^{\text{rest}} - \widehat{n\text{AUC}}_{g_2}^{\text{rest}}}{\sqrt{\text{Var}\left(\widehat{n\text{AUC}}_{g_2}^{\text{rest}} - \widehat{n\text{AUC}}_{g_2}^{\text{rest}}\right)}} \sim \mathcal{N}(0,1)$$

4 Simulations et Résultats

Nous avons testé notre méthode en l'appliquant sur des données simulées à partir d'un modèle à effets mixtes impliquant 2 groupes de vaccination où les effets fixes spécifiques aux groupes et les effets aléatoires sont modélisés par des fonctions B-splines cubiques comme décrit ci-dessous.

$$Y_{ij,g} = \gamma_0 + \mathbf{1}_{[g=1]} \sum_{k=1}^{K_1} \beta_k^1 \phi_k^1(t_{ij,1}) + \mathbf{1}_{[g=2]} \sum_{k=1}^{K_2} \beta_k^2 \phi_k^2(t_{ij,2}) + b_{0i} + \sum_{k=1}^{K_i} b_{ki} \Psi_k^i(t_{ij,g}) + \varepsilon_{ij}$$

où K_g et K_i sont le nombre de fonctions de bases des courbes splines des effets fixes et aléatoires, les ϕ_k^g et Ψ_k^i sont les $k$ème fonctions de bases de splines respectivement des effets fixes et aléatoires avec β_k^g et b_{ki} leurs coefficients de régression respectifs.

Pour nos simulations, le nombre de noeuds internes des bases de splines à l'échelle des groupes (ϕ_k^g) et individuelle (Ψ_k^i) a été fixé à 2 menant à $K_g = K_i = 5$, $\forall g \in \{1, 2\}$ et $i \in \{1, \cdots, N\}$. Les positions de ces noeuds ont été fixées à $(0.25, 5.62)$ semaines pour les bases des effets fixes et à $(2.0, 4.5)$ semaines pour les effets aléatoires. Par ailleurs, nous avons défini la matrice de variance-covariance des effets aléatoires $\mathbf{\Omega}$ comme matrice diagonale telle que $\mathbf{\Omega} = \sigma_b^2 \mathbf{I}_{K_i+1}$.
Afin de vérifier le bon comportement de notre méthode vis-à-vis des conditions de simulation et de sa capacité à gérer les données manquantes, plusieurs jeux de données ont été simulés, sous différentes conditions. Pour chaque condition de simulation, la charge virale au cours de l'ATI a été mesurée à intervalle de temps constant tel que $t = (0, 1, 2 \cdots, 24)^T$ et considérant le nombre de sujets par groupe variant entre $n_g = 20, 50$ et 100. De plus, nous avons fixé la variance des termes d'erreurs telle que $\mathbf{\Theta} = \sigma_e^2 \mathbf{I}$ où $\sigma_e^2 = 0.2$. Les différents jeux de données considérés ont été simulés de manière à faire varier la différence de nAUC entre les deux groupes de traitement telle que $\Delta n\text{AUC} = 0, -0.1$ et -0.25 $\log_{10}$ cp/ml. L'impact de la variabilité des données au sein de chaque groupe sur

la robustesse de la méthode a également été évaluée en testant la variance des nAUC égale à 0.02 et 0.1. Par ailleurs, en fixant $\gamma_0 = -0.44$ et choisissant différentes valeurs des paramètres de populations $(\boldsymbol{\beta^1}, \boldsymbol{\beta^2})$ nous a permis de faire varier ΔnAUC. Ainsi, fixer $\boldsymbol{\beta^1} = (-0.55, 4.72, 4.96, 5.18, 4.64)$ pour toutes les simulations et $\boldsymbol{\beta^2} = \boldsymbol{\beta^1}$, $(-0.54, 4.61, 4.85, 5.07, 4.54)$ et $(-0.52, 4.46, 4.69, 4.90, 4.39)$ nous a permis de cibler les différentes valeurs de ΔnAUC respectivement souhaitées.
Pour chaque combinaison de n_g, ΔnAUC et Var(nAUC) nous avons testé la méthode en considérant les données complètes, les données censurées à gauche par la présence d'une limite de détection (LOD) fixée à 50 cp/ml, ainsi que les données MAR monotones. Ces données manquantes ont été générées telle que pour tout sujet i au temps j, la variable $Y_{ij,g}$ est considérée comme manquante si $Y_{ij,g} \in \{Y_{ij,g} \mid \exists j^{'} \leqslant j, \{Y_{ij^{'},g} \geqslant \alpha\} \cap \{Y_{ij^{'}-1,g} \geqslant \alpha\}\}$, où α représente le seuil fixe de sortie d'étude. En terme plus littérale, un patient est considéré comme exclu définitivement de l'étude si son niveau de charge virale excède le seuil α au cours de deux mesures consécutives. Trois valeurs du seuil α ont été testés : $\alpha = 100.000, 50.000$ et 10.000 cp/ml (equiv. $5, 4.7$ et $4\log_{10}$ cp/ml). La considération de ces trois valeurs de seuil a permis en particulier d'évaluer notre méthode pour des pourcentages de patients quittant l'étude allant de 5 à 100% en fonction des conditions de simulations. Contrairement aux données manquantes monotones traitées comme données non disponibles (NA) n'impactant pas littéralement l'estimation du modèle à effets mixtes, l'approximation des paramètres de ce dernier en présence de données censurées à gauche par LOD requiert l'utilisation de méthodes déjà développées incluant la probabilité de données censurées dans le calcul de la vraisemblance ([3, 4, 5, 6]). A ces fins, nous avons utilisé le package R *lmec* [7] pour estimer nos modèles.

La robustesse de la méthode à estimer la différence d'aire sous la courbe normalisée par le temps de suivi a été évaluée au moyen des erreurs de Type-I et des puissances ainsi que par l'estimation du biais de ΔnAUC et de son erreur standard. Par la suite, nous avons comparé ces grandeurs obtenus par notre méthode avec celles obtenues par des méthodes adhoc se basant sur les estimations individuelles des nAUC telles que la méthode LOCF imputant les données manquantes à la dernière valeur connues ou la méthode d'imputation à la moyenne. Nous avons également comparé les résultats de notre test paramétrique avec ceux obtenus par le test non-paramétrique développé par Vardi et al. [8] correspondant à un test par permutation à l'échelle individuelle avec calcul d'AUC sur des temps restreints de suivi tels que défini dans notre méthode.

La comparaison des résultats entre les différentes méthodes montre des erreurs de Type-I équivalentes et en adéquation avec les valeurs nominales attendues pour toutes les méthodes et pour toutes les conditions de simulation en l'absence de censure de suivi. La considération de données censurées par sortie d'étude montre les limites des méthodes adhoc avec une inflation de leur erreur de Type-I sous certaines conditions. Seul un pourcentage de censure de suivi supérieur à 50% permet de différencier notre méthode du test non-paramétrique qui présente alors des résultats plus robustes que ce dernier. Ces

résultats sont confirmés par des valeurs de biais et d'erreurs standard de ΔnAUC plus faibles dans le cas de notre méthode.

Afin de montrer l'applicabilité de la méthode proposée sur des données réelles, nous l'avons utilisée pour évaluer l'efficacité de vaccins testé au cours de deux essais cliniques de vaccination thérapeutique contre le VIH. Les résultats obtenus montre alors la supériorité de notre méthode vis-à-vis des méthodes adhoc et du test non-paramétrique à détecter une différence d'AUC existante.

Bibliographie

[1] John Spritzler, Victor G DeGruttola, and Lixia Pei. Two-sample tests of area-under-the-curve in the presence of missing data. *The international journal of biostatistics*, 4(1), 2008.

[2] Melanie L Bell, Madeleine T King, and Diane L Fairclough. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data : summary measures versus summary statistics. *SAGE Open*, 4(2) :2158244014534858, 2014.

[3] Hélène Jacqmin-Gadda, Rodolphe Thiébaut, Geneviève Chêne, and Daniel Commenges. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 1(4) :355–368, 2000.

[4] Rameela Chandrasekhar, Yi Shi, Alan D Hutson, and Gregory E Wilding. Likelihood-based inferences about the mean area under a longitudinal curve in the presence of observations subject to limits of detection. *Pharmaceutical Statistics*, 14(3) :252–261, 2015.

[5] Larissa A Matos, Marcos O Prates, Ming-Hui Chen, and Victor H Lachos. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, pages 1323–1345, 2013.

[6] Florin Vaida and Lin Liu. Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, 18(4) :797–817, 2009.

[7] Florin Vaida and Lin Liu. *lmec : Linear Mixed-Effects Models with Censored Responses*, 2012. R package version 1.0.

[8] Yehuda Vardi, Zhiliang Ying, and Cun-Hui Zhang. Two-sample tests for growth curves under dependent right censoring. *Biometrika*, 88(4) :949–960, 2001.

Uncertain trees : dealing with uncertain inputs in regression trees. Application to a welding procedure qualification

Sami Alkhoury, Emilie Devijver, **Marianne Clausel**, Myriam Tami,
Eric Gaussier, Georges Oppenheim

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
sami.alkhoury@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
emilie.devijver@univ-grenoble-alpes.fr
Université de Lorraine, CNRS, IECL, Nancy, France
marianne.clausel@univ-lorraine.fr
Univ. Paris-Saclay, CentraleSupélec, MICS, Gif-sur-Yvette, France
myriam.tami@centralesupelec.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
eric.gaussier@imag.fr
Univ. Paris-Est-Marne la Vallée, Département de Mathématiques, Marne-la-Vallée, France
georges.oppenheim@gmail.com

Abstract. The goal of this work is to replace this experimental step-by-step improvement process by a numerical one based on a predictive model of weld's quality. Our contribution is to show that we can take benefit of this intrinsic variability to develop a valuable prediction model overperforming classical ones.
Our approach, referred to as uncertain trees, consists in introducing a generalization of standard regression trees introduced in (Breiman et al., 1984) dealing with uncertain input variables. One considers that an observation, even though it belongs to a given physical region, can still be associated to any region with a certain weight that depends on the distance between the observation and the statistical region. We prove that the method is theoretically well grounded, in particular regarding the consistency of uncertain trees (convergence of the approximation towards the unknown function to predict), which we establish extending results already known in the classical context (Györfi et al., 2002; Scornet et al., 2015).
Experiments conducted on classical data sets illustrate the good behavior of uncertain trees with respect to other classical approaches as standard decision trees or soft decision trees (Irsoy et al., 2012). We make use here of 12 data sets, namely Abalone (AB), Ailerons (AL), Bike-Day (BD), Bike-Hour (BH), Boston (BO), Diabetes (DI), Facebook Comments (FC), Forest Fires (FF), Ozone (OZ), Skill (SK), Super Conductor (SC) and Video Transcoding (VT), all commonly used in regression tasks.

We also present the performances of uncertain regression trees in the case of the Ultimate Tensile Strength variable prediction. Uncertain trees significantly outperform both standard and soft trees, on almost all data sets.
Finally we intend to develop an uncertain version of Gradient Boosted Trees to learning welding mechanical prediction with both quantitative and qualitative variables. The question of adding uncertainty output analysis as in (Meinshausen, 2006; Zheng, 2012) will also be discussed.

Keywords. Welding quality; Uncertain inputs; Ensemble methods.

On the approximation of extreme quantiles with neural networks

Michaël Allouche [1] , Stéphane Girard [2] & Emmanuel Gobet [3]

[1] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France; michael.allouche@polytechnique.edu.*

[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; stephane.girard@inria.fr.*

[3] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France; emmanuel.gobet@polytechnique.edu.*

Résumé. Dans cette étude nous proposons une nouvelle paramétrisation du générateur d'un réseau antagoniste génératif (GAN) adaptée aux données issues d'une distribution à queue lourde. Nous apportons une analyse de l'erreur d'approximation en norme uniforme d'un quantile extrême par le GAN ainsi construit. Des simulations numériques sont réalisées sur des données réelles et simulées.

Mots-clés. Théorie des valeurs extrêmes, réseau de neurones, modèle génératif

Abstract. In this study, we propose a new parametrization for the generator of a Generative adversarial network (GAN) adapted to data from heavy-tailed distributions. We provide an analysis of the uniform error between an extreme quantile and its GAN approximation. Numerical experiments are conducted both on real and simulated data.

Keywords. Extreme value theory, neural networks, generative models

1 Introduction

In this paper we are interested in approximating the quantile function q_X defined by $q_X(u) := F_X^{\leftarrow}(u) = \inf\{x : F_X(x) \geq u\}$, for all level of quantiles $u \in [0, 1)$, where F_X is an unknown cumulative distribution function on $\mathcal{X} \subseteq \mathbb{R}$. Clearly, extreme quantiles are observed as $u \to 1$ and will be our region of interest. The objective is, starting from an i.i.d. sample set $\{X_i \in \mathcal{X}\}_{i=1}^n$, to build a generator $G_\theta : \mathcal{Z} \to \mathcal{X}$ from a parametric family of functions $\mathcal{G} = \{G_\theta\}_{\theta \in \Theta}$ and mapping a random variable $Z : \mathcal{Z} \to \mathbb{R}^d$ with known density p_Z to the data support $\mathcal{X}$. In this study, we shall consider $Z \sim \mathcal{U}(0, \mathbf{I}_d)$ and denote by p_θ the density associated with $G_\theta(Z)$. Then, for each p_θ, $\theta \in \Theta$, is a potential candidate to approximate $p_X = F_X'$. In a neural network architecture, this setting is related to Generative Adversarial Networks (GAN) [7].

Let X be a random variable associated with F_X supposed to be continuous and strictly increasing. We focus on the case of heavy-tailed distributions, *i.e.* when F_X is attracted to the maximum domain of Pareto-type distributions with tail-index $\gamma > 0$. From [2], the survival function $\bar{F}_X := 1 - F_X$ of such a heavy-tailed distribution can be expressed as

($\mathbf{H_1}$): $\bar{F}_X(x) = x^{-1/\gamma}\ell_X(x)$, where ℓ_X is a slowly-varying function at infinity *i.e.* such that $\ell_X(\lambda x)/\ell_X(x) \to 1$ as $x \to \infty$ for all $\lambda > 0$.

In such a case, $\bar{F}_X$ is said to be regularly-varying with index $-1/\gamma$ at infinity, which is denoted for short by $\bar{F}_X \in RV_{-1/\gamma}$. The tail-index γ tunes the tail heaviness of the distribution function F_X. Assumption $\mathbf{(H_1)}$ is recurrent in risk assessment, since actuarial and financial data are most of the time heavy-tailed, see for instance the recent studies [1, 3] or the monographs [6, 9]. As a consequence of the above assumptions, the tail quantile function $x \mapsto q_X(1 - 1/x)$ is regularly-varying with index γ at infinity, see [5, Proposition B.1.9.9], or, equivalently,

$$q_X(u) = (1-u)^{-\gamma} L\left(\frac{1}{1-u}\right), \tag{1}$$

for all $u \in (0,1)$ with L a slowly-varying function at infinity. Clearly $q_X(u) \to \infty$ as $u \to 1$ but is pretty smooth elsewhere. This type of function does not seem to be consistent with a neural network approximation framework since the latter mainly consists in making a linear combination of bounded functions, which is very unlikely to approximate diverging functions. This argument is confirmed by the Universal Approximation Theorem [4] stating that a one hidden layer neural network can approximate any continuous function on a **compact set**.

2 Contribution

In order to build a tail-index function which may be well approximated by a neural network, the quantile function (1) is rewritten in a logarithmic scale and normalized to avoid exploding issues at the boundaries. Without of loss of generality, one can assume that $\eta := \mathbb{P}(X \geq 1) \neq 0$ and, since, we focus on the upper tail behavior of X, introduce the random variable $Y = X$ given $X \geq 1$. It follows that the quantile function of Y is given by $q_Y(u) = q_X(1 - (1-u)\eta)$, for all $u \in (0,1)$. Thus, we define the Tail-Index function (TIF) as

$$f^{\mathrm{TIF}}(u) := -\frac{\log q_X\big(1-(1-u)\eta\big)}{\log(1-u^2) - \log 2},$$

for all $u \in (0,1)$. The main contribution of this work is to combine the TIF analysis based on Extreme Value Theory with GANs in order to address the general issue of

neural network approximation for extremes. Let $\varphi^{\mathrm{TIF}} : \mathbb{R}^2 \to \mathbb{R}$ be the non-linear TIF transformation with

$$\varphi^{\mathrm{TIF}}(x,u) := \left(\frac{1-u^2}{2}\right)^{-x}.$$

In addition, let $\mathbf{e} \in \mathbb{R}^6$ be a vector of functions mapping from $[0,1]$ to $\mathbb{R}$ and $\alpha \in \mathbb{R}^6$ be a vector of parameters. Thus, we define the **Tail-GAN** with a generator $G_\psi^{\mathrm{TIF}} : \mathbb{R}^d \to \mathbb{R}$ where the j-th output component is defined as

$$G_\psi^{\mathrm{TIF(j)}}(Z) = \varphi^{\mathrm{TIF}}\left(G_\psi^{(j)}(Z), Z^{(j)}\right),$$

and $\psi = (\theta, \alpha)$ with

$$G_\psi^{(j)}(Z) = G_\theta^{(j)}(Z) + \left\langle \boldsymbol{e}\left(Z^{(j)}\right), \alpha \right\rangle.$$

The selection of functions in $\boldsymbol{e}$ and optimal parameter α is based on the following approximation results dealing with the regularity properties of f^{TIF} and the construction of its regularized extension.

3 Approximation results

Our first result describes the behaviour of f^{TIF} in the neighborhood of 0 and 1.

Proposition 1 *Under* $(\mathbf{H}_1)$, f^{TIF} *is a continuous and bounded function on* $[0,1]$. *Besides,* $f^{\mathrm{TIF}}(0) = 0$ *and* $f^{\mathrm{TIF}}(u) \to \gamma$ *as* $u \to 1$.

Focusing on the behavior of the first derivative of the TIF, extra assumptions on F_X, or equivalently on L, are necessary such that f^{TIF} is differentiable. Consider the Karamata representation of the slowly-varying function L [5, Definition B1.6]:

$$L(x) = c(x) \exp\left(\int_1^x \frac{\varepsilon(t)}{t}\,\mathrm{d}t\right),$$

where $c(x) \to c_\infty$ as $x \to \infty$ and ε is a measurable function such that $\varepsilon(x) \to 0$ as $x \to \infty$. Our second main assumption then writes:

$(\mathbf{H_2})$: $c(x) = c_\infty > 0$ for all $x \geq 1$ and $\varepsilon(x) = x^\rho \ell(x)$ with $\ell \in RV_0$ and $\rho < 0$.

The assumption that c is a constant function is equivalent to assuming that L is normalized [8] and ensures that L is differentiable. The condition $\varepsilon \in RV_\rho$ with $\rho < 0$ entails that $L(x) \to L_\infty \in (0,\infty)$ as $x \to \infty$. Besides, $(\mathbf{H_2})$ entails that F_X satisfies the so-called second-order condition which is the cornerstone of all proofs of asymptotic normality in extreme-value statistics. We shall also consider the assumption:

($\mathbf{H_3}$): ℓ is normalized.

The latter condition ensures that ℓ is differentiable on $(0,1)$ and thus that L and q_X are twice differentiable on $(0,1)$. Our second result provides the behaviour of the first order derivative of f^{TIF} in the neighborhood of 0 and 1.

Proposition 2 *Assume* ($\mathbf{H}_1$) *and* ($\mathbf{H}_2$) *hold. Then,* f^{TIF} *is continuously differentiable on* $(0,1)$ *and*

$$\begin{aligned} \partial_u f^{\mathrm{TIF}}(0) &= \frac{\gamma + \varepsilon\,(1/\eta)}{\log(2)}, \\ \partial_u f^{\mathrm{TIF}}(u) &\to \infty \ \textit{as}\ u \to 1. \end{aligned} \tag{2}$$

It is possible to build regularized version of f^{TIF} by removing the diverging components in the neighborhood of $u = 1$. To this end, consider

$$f^{\mathrm{R}}(u) := f^{\mathrm{TIF}}(u) - \langle \boldsymbol{e}(u),\, \alpha \rangle\,, \tag{3}$$

where $\boldsymbol{e} : \mathbb{R} \to \mathbb{R}^6$ is not described here for the sake of conciseness. Regularity properties of f^{R} are established in the next Proposition.

Proposition 3

(i) If ($\mathbf{H_1}$) *holds, then*

$$\lim_{u \to 0} f^{\mathrm{R}}(u) = \lim_{u \to 1} f^{\mathrm{R}}(u) = 0. \tag{4}$$

(ii) If, moreover, ($\mathbf{H_2}$) *holds with* $\rho < -1$*, then* f^{R} *is continuously differentiable on* $[0,1]$ *and*

$$\lim_{u \to 0} \partial_u f^{\mathrm{R}}(u) = \lim_{u \to 1} \partial_u f^{\mathrm{R}}(u) = 0. \tag{5}$$

(iii) If, moreover, ($\mathbf{H_3}$) *holds with* $\rho < -2$*, then* f^{R} *is twice continuously differentiable on* $[0,1]$.

Given the above regularity properties, it is possible to establish the convergence rate of the uniform error for a one hidden layer neural network depending on the parameter ρ.

Theorem 4 *Let* σ *be a ReLU function. There exists a neural network with* J *neurons and real coefficients* $\{\gamma_j, \lambda_j, b_j\}_{j=1,\dots,J}$ *such that:*

1. *For* $-2 < \rho < -1$,

$$\sup_{t \in [0,1]} \left| f^{\mathrm{R}}(t) - \sum_{j=1}^{J} \gamma_j \sigma\left(\lambda_j t + b_j\right) \right| = \mathcal{O}\left(J^{\rho}\right),$$

2. for $\rho \leq -2$,

$$\sup_{t\in[0,1]} \left| f^{\mathrm{R}}(t) - \sum_{j=1}^{J} \gamma_j \sigma\left(\lambda_j t + b_j\right) \right| = \mathcal{O}\left(J^{-2}\right).$$

Numerical experiments will be presented in the communication in order to compare the realizations of traditional GANs with our Tail-GAN in two situations: simulated data from heavy-tailed distributions and real financial data from public source.

References

[1] J. Alm. Signs of dependence and heavy tails in non-life insurance data. *Scandinavian Actuarial Journal*, 2016(10):859–875, 2016.

[2] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987.

[3] V. Chavez-Demoulin, P. Embrechts, and S. Sardy. Extreme-quantile tracking for financial time series. *Journal of Econometrics*, 181(1):44–52, 2014.

[4] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4):303–314, 1989.

[5] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

[6] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, Bing X., D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] E. Kohlbecker. Weak asymptotic properties of partitions. *Transactions of The American Mathematical Society*, 88(2):346–365, 1958.

[9] S. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

A generalized method for Sparse Partial Least Squares (Dual-SPLS): theory and applications

Louna Alsouki[1,3], François Wahl[1,2], Laurent Duval[2], Clément Marteau[1] & Rami El-Haddad[3]

[1] *Université Claude-Bernard Lyon 1, 43 boulevard du 11 Novembre 1918, 69100 Villeurbanne, France,*
[2] *IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France,*
[3] *Université Saint Joseph de Beyrouth, Mar Roukoz – Dekwaneh, B.P. 1514, Liban,*
louna.al-souki@univ-lyon1.fr, francois.wahl@math.univ-lyon1.fr, laurent.duval@ifpen.fr, marteau@univ-lyon1.fr, rami.haddad@usj.edu.lb

Résumé. En analyse des données, la grande dimensionalité est souvent un obstacle délicat à surmonter qui être résolu en représentant les données dans un espace de dimension inférieure en utilisant des méthodes de projection comme la régression des moindres carrés partiels (PLS) [1] ou en ayant recours à des méthodes de sélection de variables comme l'approche lasso [2]. La Sparse Partial Least Squares (SPLS) vise à résoudre le problème d'interpretation des coefficients en combinant les deux procédures. Plusieurs implémentations ont été proposées [3, 4, 5]. Cependant, des problèmes de précision de prédictions et de bonne interprétation des coefficients surgissent dans ces travaux. C'est pourquoi nous avons développé la Dual Sparse Partial Least Squares, une méthode flexible qui permet d'obtenir des prédictions plus précises et une meilleure interprétation des coefficients grâce à leur parcimonie. Dans cet article, nous présentons la théorie derrière Dual-SPLS et certains résultats d'applications sur des ensembles de données pétrolières réelles.

Mots-clés. Moindres carrés partiels, parcimonie, régression, norme duale, algorithme lasso.

Abstract. In data analysis, high dimensionality is often a delicate obstacle to overcome which can be solved by representing the data in a lower dimensional space using projection methods like the Partial Least Squares regression (PLS) [1] or by resorting to variable selection methods like the lasso approach [2]. The Sparse Partial Least Squares (SPLS) aims at solving the interpretation problem of the coefficients by combining the two latter. Several implementations have been proposed [3, 4, 5]. However, problems of accuracy of predictions and correct interpretation of regression coefficients arise in these approaches. Hence we developed the Dual Sparse Partial Least Squares, a flexible method that results in more accurate predictions and better interpretation of the coefficients due to their sparsity. In this paper we present the theory behind Dual-SPLS and some applicative results on petroleum data sets.

Keywords. Partial Least Squares, sparsity, regression, dual norm, lasso algorithm.

1 Introduction

Regression analysis helps in inferring relationships between data sets, with the additional objective of extracting interpretable information. However, a recurrent problem haunting statistical data analysis is data high dimensionality. One can choose to tackle this issue by using dimension reduction methods, like the PLS procedure [1], allowing to represent the data in a lower dimensional space. It reduces the dimensionality by selecting derived components. It is an iterative method that deals with highly correlated data and results in accurate outcomes. Algorithms are generally straightforward and simple to handle without matrix inversion. However, regression coefficients are frequently hard to interpret (see section 3). Another suggestion often considered is variable selection, like in the lasso [2]. It performs regularization in order to enhance the prediction accuracy, while simplifying the interpretation of the regression coefficients due to the sparsity of the representation. Nevertheless, the lasso is very sensitive to the type of data and does not always result in interpretable coefficients: in fact, it selects at most n variables before it saturates [6]. Sparse Partial Least Squares (SPLS) [3, 4, 5] combines both approaches by adding to the PLS framework a selection step inspired by the lasso. It is represented by the following optimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\{-\hat{\mathrm{Cov}}(\mathbf{X}\mathbf{w}, \mathbf{y}) + \lambda_s \|\mathbf{w}\|_1\}, \qquad \text{for } \mathbf{w}^T\mathbf{w} = 1, \tag{1}$$

under the orthogonality constraint of components, with sparsity parameter $\lambda_s > 0$.

Lê Cao *et al.* (2008) [3] and Chun and Keleş (2010) [4] developed SPLS approaches that both give an approximate solution. Thus, Durif *et al.* [5] conceived a similar method in the context of classification that solves exactly Problem (1) in the univariate response case. Nonetheless, it can be applied in the regression. It however appears to be time consuming on high dimensional data.
Inspired by these methodologies, we devised a new strategy called Dual Sparse Partial Least Squares (Dual-SPLS) that provides prediction accuracy equivalent to the PLS method along with easier interpretation of regression coefficients thanks to the sparsity of the results. Moreover, it generalizes the above mentioned approaches on the theoretical point of view.

We first present the main ingredients of the Dual-SPLS. We then show some results of applications on petroleum data sets.

2 Dual Sparse Partial Least Squares

The proposed method originated from noticing the similarity between the variational formulation of the PLS (with the PLS1 methodology) and the expression of the dual norm of a vector.

Let $\Omega(\cdot)$ be a norm on $\mathbb{R}^P$. The associated dual norm, denoted $\Omega^*(\cdot)$, is defined, for any $\mathbf{z} \in \mathbb{R}^P$, as:

$$\Omega^*(\mathbf{z}) = \max_{\mathbf{w}}(\mathbf{z}^T\mathbf{w}) \quad \text{s.t.} \quad \Omega(\mathbf{w}) = 1. \tag{2}$$

Meanwhile, the optimization problem solved by the PLS method for the first component writes:

$$\max_{\mathbf{w}}(\mathbf{y}^T\mathbf{X}\mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_2 = 1. \tag{3}$$

Comparing (2) and (3), one notices that optimizing the PLS criterion amounts to finding the vector $\mathbf{w}_1$ that goes with the conjugate of the ℓ_2-norm of $\mathbf{z}$ where $\mathbf{z} = \mathbf{X}^T\mathbf{y}$, which can be exploited in evaluating different norm expressions.
As in the lasso approach, we consider the combination of the ℓ_1 and ℓ_2 norms:

$$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2 \,. \tag{4}$$

The closed form solution can be expressed with the soft thresholding operator. Since we are dealing with a vector, we can consider each coordinate $p \in \{1, \ldots, P\}$ of $\mathbf{w}$ and the solution can be written as:

$$\frac{w_p}{\|\mathbf{w}\|_2} = \frac{1}{\mu}\delta_p(|z_p| - \nu)_+ \tag{5}$$

where $\boldsymbol{\delta}$ is the vector of the signs of $\mathbf{z}$, μ guarantees the normality constraint in (2) and $\nu = \lambda\mu$. This is relevant since we can compare ν to z_p, and therefore shrink to zero the coefficients that correspond to the small coordinates of $\mathbf{z}$ (compared to ν), which enforces sparse regression coefficients.

However, the main challenge resides in setting the parameter ν, which affects the amount of shrinkage. We propose to choose it iteratively and adaptively according to the number of variables that we would like to keep in the active set at each iteration. In other words, for each number i of desired components, an optimal ν_i is chosen to impose a given proportion of null coefficients. The Dual-SPLS method is implemented in the form of Algorithm 1.

Algorithm 1 DUAL-SPLS ALGORITHM FOR $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2$

Input: $\mathbf{X}_1, \mathbf{y}, I$ (number of components)
for $i = 1, \ldots, I$ **do**
  $\mathbf{z}_i = \mathbf{X}_i^T\mathbf{y}$ (weight vector)
  Find ν in the adaptive way
  $\mathbf{z}_\nu = (\delta_p(|z_p| - \nu)_+)_{p\in}$ (applying the threshold)
  $\mu = \|\mathbf{z}_\nu\|_2 \quad \lambda = \dfrac{\nu}{\mu} \quad \mathbf{w}_i = \dfrac{\mu}{\nu\|\mathbf{z}_\nu\|_1 + \|\mathbf{z}_\nu\|_2^2}\mathbf{z}_\nu$ (loadings)
  $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i / \|\mathbf{X}_i\mathbf{w}_i\|$ (scores)
  $\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i^T\mathbf{t}_i\mathbf{X}_i$ (deflation)
end for
Compute regression coefficients

3 Results and discussions

3.1 Data sets

The data set is composed of 243 NMR spectra of refined oil samples. Each spectrum is originally represented by more than 65000 variables. However, we have pretreated them by eliminating irrelevant parts, removing repeated observations and normalizing amplitudes between 0 and 1, which leaves us with around 21000 variables and 182 observations. Our aim is to predict the density of these oil samples.

3.2 Benchmark

We assess the efficiency of the Dual-SPLS by computing the root mean square error (RMSE) for prediction performance and then we examine the interpretation of the coefficients by comparing them to the original raw spectra. The evaluation is organized as a benchmark comparing the following methods together: PLS [1], sPLS of Lê Cao *et. al.* (as implemented in mixOmics) [3], SPLS of Keleş *et. al.* (as implemented in spls) [4], SPLS of Durif *et. al.* (as implemented in plsgenomics) [5] and lasso in glmnet package [2]. In Figure 1, the calibration and validation sets are chosen adequately. The figure is divided into two parts: the left part corresponds to the RMSE values of the validation set according to the number of components and the right part represents the coefficients of each regression. For PLS related methods we select 6 components. As for Figure 2, the calibration and validation sets are chosen randomly. We applied the regression methods for 100 repetition in order to represent the boxplots and compare the results.

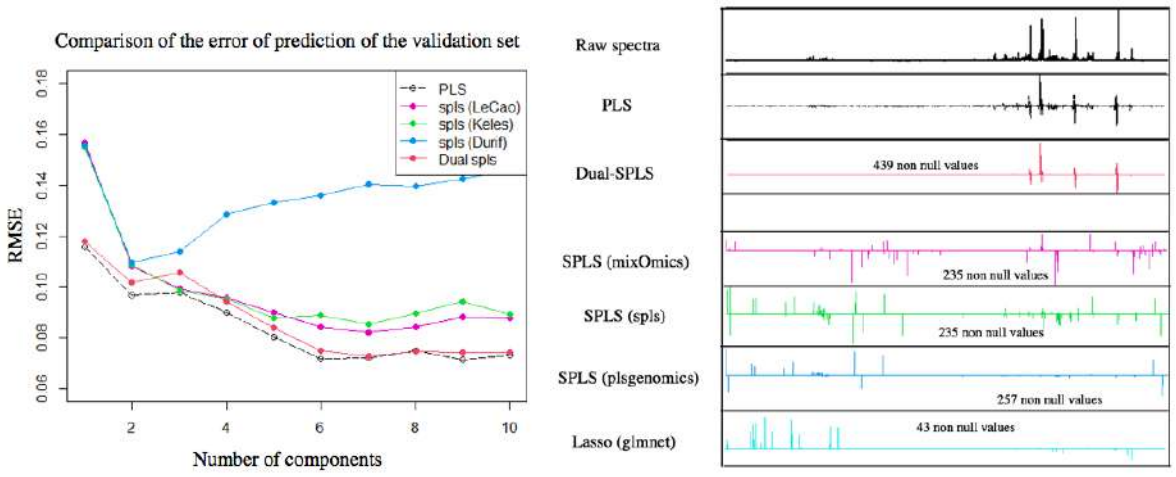


Figure 1: Benchmark of (sparse) PLS methods on the NMR data set: prediction error according to the number of component (left), raw data and coefficients localization (right).

In the Figures 1 and 2 , we require a 99 % proportion of null coefficients while applying the Dual-SPLS and use cross validation to choose the adequate amount of penalization λ_s for each of the other cases. Note that the x-axis is not represented with chemically-sound units due to preprocessing.

From Figure 1 (left), all methods almost match the prediction accuracy of the PLS from two components on, except for spls from the plsgenomics package [5] whose predictions are slightly less accurate. The lasso algorithm provides RMSE values around 0.09 according to the choice of shrinkage parameter. We even notice that the closest results to the PLS are those from the new approach. To compare coefficients localization, we select six components for PLS-related methods as the RMSE curves tend to plateau above this value. On Figure 1 (right), we see similarities between the PLS coefficients and the raw spectra, however, no information concerning the localization of the most relevant variables can be extracted. The Dual-SPLS solves this problem with sparse results that selects the most important variables. As for the rest of the methods, the results are misleading and does not properly indicate which variable are best to keep.

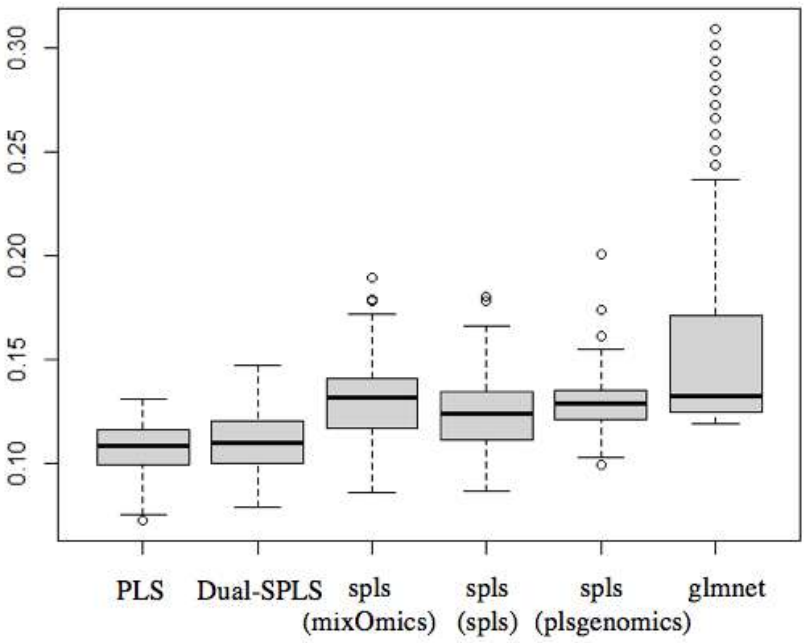


Figure 2: Benchmark of (sparse) PLS methods on the NMR data set: prediction error boxplots using random calibration.

From Figure 2, where we compare the boxplots of the methods applied for six components (for PLS-related methods), we conclude the same: the prediction accuracy of the PLS is the most similar by using the Dual-SPLS.

4 Conclusions

The Dual-SPLS introduces a general framework providing a novel family of regression methods that encompasses the standard PLS method. It offers the possibility to use a quantity of different norm shapes. In the case of a norm inspired by the lasso, it already preserves the prediction accuracy of the PLS and previously proposed sparse PLS methodologies. On NMR data it shows to be even sparser with better localized and more interpretable coefficients. The next steps will consist first in implementing and sharing this method as an R package, and second in evaluating the gain in performance by using other norms mimicking the fused or the group lasso.

References

[1] H. Wold, "Path models with latent variables: The NIPALS approach," in *Quantitative Sociology. International Perspectives on Mathematical and Statistical Modeling*, pp. 307–357. Elsevier, 1975.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[3] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Stat. Appl. Genet. Mol. Biol.*, vol. 7, no. 1, pp. 35, 2008.

[4] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 72, no. 1, pp. 3–25, 2010.

[5] G. Durif, L. Modolo, J. Michaelsson, J. E. Mold, S. Lambert-Lacroix, and F. Picard, "High dimensional classification with combined adaptive sparse PLS and logistic regression," *Bioinformatics*, vol. 34, no. 3, pp. 485–493, Feb. 2018.

[6] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

Spatial sampling and spatial entropy

Linda Altieri and Daniela Cocchi

Keywords: Environmental sampling, spatially correlated Poisson sampling, sampling entropy

1 Introduction

The objective of spatial sampling is to collect samples, i.e. subsets of individuals from a population, in the 2-dimensional space, in order to estimate some population characteristics. Spatial sampling is strongly linked to environmental sampling. Such expression states that the data spatial location is a fundamental information, and that sampling techniques for environmental data are mutuated from the theory of spatial sampling. Under the viewpoint of the reference population, environmental sampling focuses on natural populations. Examples can be found in biology, geography, landscape studies, forestry, and in the study of environmental dangers such as wildfires, earthquakes, polluting agents.

In finite population inference, the design-based context aims at estimating population quantities, considered as unknown but fixed. In this case the only source of randomness is the probability of the samples, which is related to the inclusion/extraction probabilities of each population element. Information may be available for moving such individual probabilities from equality. In particular, information related to space may be organized in this respect.

The link between sampling and entropy has been extensively debated in statistics (Shewry and Wynn, 1987; Lee, 2006). The search for sampling plans with high entropy is an important task in survey sampling design-based theory. Sample selection should follow the idea of randomization: a sampling design should assign a non-null probability to as many samples as possible. A widely accepted measure of randomness of a sampling design is its entropy (Tillé and Haziza, 2010; Tillé and Wilhelm, 2017): a sampling design has high entropy when there is a high amount of uncertainty or surprise in the sample to select. Conditional Poisson sampling has been identified as the maximum entropy sampling design when the sample size is fixed (Hajek, 1981; Tillé, 2006; Tillé and Wilhelm, 2017). Maximum entropy sampling has been deepened in computer science and received important contributions in such field (Ko et al., 1995).

Under a different perspective, entropy is a popular heterogeneity measure since a long time, with reference to any kind of random variables. After being firstly introduced in information theory (Shannon, 1948), it rapidly became popular in many applied sciences to measure the degree of heterogeneity among observations. In its original proposal, entropy does not take space into account. A rather recent research field aims at accounting for space in entropy measures. In this spirit, a sequel of papers (Altieri et al., 2018a, 2019a,b) exploits the

decomposition of bivariate distributions linked to entropy in order to quantify the contribution of spatial association to the entropy of a variable. Euclidean distances between spatial locations are employed for constructing the second variable. Such spatial entropy measures are employed in this exposition to improve a sequential spatial design.

In what follows we refer to the basic concept above as "spatial entropy", while the entropy of the sampling design is to be named, rather, sampling entropy. The two entropies need to be distinguished, as they refer to different aspects of the data. Spatial entropy refers to the spatial correlation of the study variable. Sampling entropy refers to the randomness of the potential samples; it regards the chances of selecting population units, irrespective of the value they possess for the variable under study.

The simulation study and all computations are implemented via the `R` software, with the help of the packages `SpatEntropy` (Altieri et al., 2018b) and `BalancedSampling` (Grafström and Lisic, 2018).

2 Recalling some theory

Spatially Correlated Poisson Sampling is a sequential (adaptive) technique that modifies the initial first order inclusion probabilities of the elements of a finite population according to the scheme described as follows.

Starting from the first population unit, for which a Bernoulli draw with probability π_1 is proposed, after the draw an indicator function is $I_1 = 1$ if the unit is sampled, and 0 otherwise. Then, at the general step $k = 2, \ldots, N$, the values for $I_1, \ldots, I_{k-1}$ are known and unit k is sampled with probability $\pi_k^{(k-1)}$, i.e. with an inclusion probability that was updated at the previous step, when sampling unit $k-1$. The inclusion probabilities for all remaining units $l = k+1, \ldots, N$ are updated:

$$\pi_l^{(k)} = \pi_l^{(k-1)} - (I_k - \pi_k^{(k-1)}) b_k^{(l)}. \tag{1}$$

This way, at each step k the inclusion probabilities of the visited units $1, \ldots, k$ leave the room to the corresponding indicator functions. At step N, the vector becomes $\pi_1^{(N)}, \ldots, \pi_N^{(N)} = I_1, \ldots, I_N$, which indeed sums to n.

Stressing on space, a geographical distance between population units is introduced for evaluating the component $b_k^{(l)}$, a factor that influences their selection in the sample. Negative correlation weights are attributed to units that are close in space in order to obtain the spatial spreading that is desirable for sampling. A "maximal weight strategy" has been proposed (Grafström, 2012), that produces samples of fixed size $n = \sum_{k \in U} \pi_k$ and is very efficient, provided that close units carry similar values.

We propose to enhance space highlighting in (1) by building a new weighting system that exploits the theory of spatial entropy (Altieri et al., 2018a, 2019a,b). For a certain transformation Z of the variable under study X, for which the entropy

$$H(Z) = E[I(p_Z)] = \sum_{i=1}^{I} p(z_i) \log \left(\frac{1}{p(z_i)} \right). \tag{2}$$

can be constructed, a system of distances summarized by the variable W is also defined. Such entropy $H(Z)$ can be decomposed as

$$H(Z) = SMI(Z, W) + H(Z)_W. \tag{3}$$

The first component of (3), called Spatial Mutual Information, is defined as

$$SMI(Z, W) = \sum_{m=1}^{M} p(w_m) SPI(Z|w_m) \tag{4}$$

where each mth component $SPI(Z|w_m)$, i.e. the Spatial Partial Information, describes the relationship of Z with a specific distance class w_m of W:

$$SPI(Z|w_m) = \sum_{r=1}^{R} p(z_r|w_m) \log \left(\frac{p(z_r|w_m)}{p(z_r)} \right). \tag{5}$$

.

In general, when $SMI(Z, W)$ is high, the value carried by a sampled unit gives us information about what to expect from its neighbouring units; the stronger the mutual information, the smaller our interest in sampling neighbouring units. A peculiar aspect of $SMI(Z, W)$ is that it can be decomposed into the partial terms $SPI(Z|w_m)$ of (5) at different distance ranges. Thanks to such decomposition, the distance ranges for the variable under study can be decided according to the problem and the corresponding $SPI(Z|w_m)$ terms chosen as contributions to weights $b_k^{(l)}$. This way, partial spatial information assumes the role of auxiliary variable for building a well founded weighting system for sampling. Each $SPI(Z|w_m)$ is always positive, and tunes sampling neighbouring units with a strength that depends on the spatial correlation of the study variable at the chosen distances.

Weights $b_k^{(l)}$ are assigned starting from one of the M SPI terms. In particular, if units k and l are in the mth distance range, then

$$b_k^{(l)} = \frac{SPI(Z|w_m)}{C} \qquad \text{for } d(k, l) \in w_m \tag{6}$$

where C is a normalizing constant so that $\sum_{l=k+1}^{N} b_k^{(l)} = 1$ for all k, i.e. the triangular weight matrix is row-standardized. The easiest solution is that the normalizing constant is just the sum of the unnormalized weights: $C = \sum_{l=k+1}^{N} \widetilde{b}_k^{(l)}$ with $\widetilde{b}_k^{(l)} = SPI(Z|w_m)$ for $d(k, l) \in w_m$.

Units are ordered according to some labelling in space. For instance, if spatial units are arranged over a grid, unit 1 can be the top-left unit, unit 2 can be at its right, or below, and so on. Different labelling orders return different updates in the inclusion probabilities of the remaining units; the method holds for any starting point and labelling criterion, as long as the distance between all pairs of units is well defined.

3 Simulation and results

In order to explore the potential improvement of spatially correlated Poisson sampling (SCPS) with the help of spatial partial information-based (SPI)

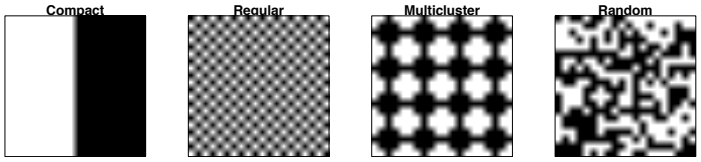


Figure 1: Basic different configurations for the same $\pi_P = 0.5$

Table 1: Spatial partial information at the four distance classes.

	$[0,1]$	$]1,2]$	$]2,5]$	$]5,20\sqrt{2}]$
Compact	0.786	0.687	0.455	0.010
Regular	0.509	0.918	<0.001	<0.001
Multicluster	0.146	0.057	0.008	<0.001
Random	0.020	0.010	0.001	<0.001

weights, we run a comparative study. Simulated binary datasets are employed to estimate the variable mean (i.e. the proportion for binary data) and to evaluate the MSE of the estimator.

Consider $N = 400$ realizations of a binary variable X with half outcomes $x_0 = 0$ and half $x_1 = 1$. The true mean/proportion is $m(X) = \sum_k x_k/400 = 0.5$. Realizations are arranged over a square observation area gridded by 20×20 pixels; each pixel is assumed to be a 1×1 square. Realizations are organized according to four different spatial configurations that produce different spatial entropy values, as in Altieri et al. (2019b). The first one is the most clustered spatial distribution, named "compact",obtained by assigning x_0 values to the pixels located at the left part of the window and x_1 values to pixels located at the right part. The second one is the most "regular" spatial distribution, corresponding to a chessboard, obtained by assigning x_0 values to pixels adjacent to x_1-valued pixels, and vice versa. The third one is a "multicluster" distribution with 16 clusters, whose centroids are regularly distributed over the area; then, x_0 values are assigned to pixels surrounding the centroids and x_1 values to the remaining pixels. The last one is a "random" pattern without spatial correlation whatsoever, obtained by assigning x_0 or x_1 values to pixels via simple random sampling without replacement from the generated sequence. The four datasets are displayed in Figure 1.

Four distance classes are chosen over the observation area: $w_1 = [0,1]$, $w_2 =]1,2]$, $w_3 =]2,5]$, $w_4 =]5,20\sqrt{2}]$, where $20\sqrt{2}$ is the maximum distance over the observation window. Reasons for choosing such classes are found in the fundamentals of spatial statistics, and are discussed in Altieri et al. (2018a, 2019b). Spatial partial information terms (5) are computed, following Altieri et al. (2019b), and the resulting values are shown in Table 1, which contains important characteristics of the population.

Afterwards, 100 samples of fixed size $n = 40$ are drawn from each dataset. The initial inclusion probabilities are constant: $\pi_k = n/N$ for all units k. Then, a sample is drawn from each dataset using the sequential approach of SCPS, where the weight assigned to each pair of units comes from the SPI value in Table 1, according to the distance between units. The SPI values are rescaled

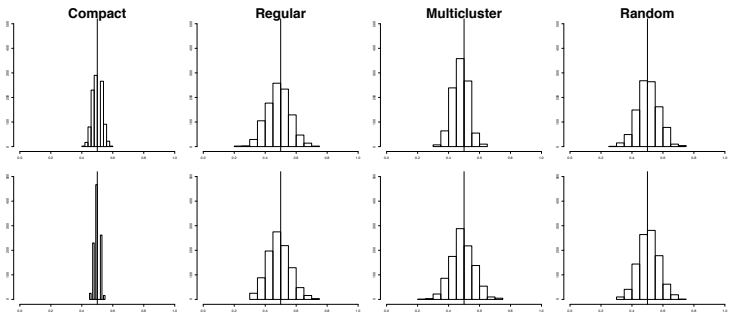


Figure 2: HT estimate for $m(X) = 0.5$; first line with SPI, second line with maximal weight

Table 2: Mean Squared Error of the HT estimates.

	Maximal weights	SPI weights
Compact	.0004	.0009
Regular	.0053	.0057
Multicluster	.0055	.0028
Random	.0049	.0052

so that they sum to 1 for each population unit, and the constraint for positive weights is checked. The SCPS is implemented, updating the remaining units' inclusion probabilities sequentially, until all population units have been either sampled or rejected and a sample size of n is reached.

All results are compared to SCPS with maximal weights as implemented by the R package `BalancedSampling` [REF]. Note that the 100 samples produced with the maximal weights system are the same across spatial configurations, since such weights only consider the distance between units and not the strength of the spatial correlation.

We compare the two weighting systems, with the sample size constrained to be n for all samples, thanks to the sequential technique and to the sum-to-1 constraint for the weights. The HT estimate for the variable mean is displayed in Figure 2, where the thick vertical line marks the true mean $m(X) = 0.5$. The MSE has been computed over the simulated samples, with results displayed in Table 2. Results with the maximal weight tevnhique are winning in the case of a compact spatial scheme. Since it is based only on the distance, a strong positive correlation between the values of the variable is hidden in this case.

References

Altieri, L., D. Cocchi, and G. Roli (2018a). A new approach to spatial entropy measures. *Environmental and Ecological Statistics 25*(1), 95–110.

Altieri, L., D. Cocchi, and G. Roli (2018b). *SpatEntropy: Spatial Entropy Measures*. R package version 0.1.0.

Altieri, L., D. Cocchi, and G. Roli (2019a). Measuring heterogeneity in urban expansion via spatial entropy. *Environmetrics 30*(2), e2548.

Altieri, L., D. Cocchi, and G. Roli (2019b). Advances in spatial entropy measures. *Stochastic Environmental Research and Risk Assessment*.

Bondesson, L. and D. Thorburn (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics 35*(3), 466–483.

Grafström, A. (2012). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference 142*(1), 139–147.

Grafström, A. and J. Lisic (2018). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.4.

Hajek, J. (1981). *Sampling from a finite population*. New York, New York: Marcel Dekker, Inc.

Ko, C.-W., J. Lee, and M. Queyranne (1995). An exact algorithm for maximum entropy sampling. *Operations Research 43*(4), 684–691.

Lee, J. (2006). Maximum entropy sampling. *Encyclopedia of Environmetrics 4*.

Shannon, C. (1948). A mathematical theory of communication. *Bell Dyditem Technical Journal 27*, 379–423, 623–656.

Shewry, M. C. and H. P. Wynn (1987). Maximum entropy sampling. *Journal of Applied Statistics 14*(2), 165–170.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Tillé, Y. and D. Haziza (2010). An interesting property of the entropy of some sampling designs. *Survey Methodology 36*(2), 229–231.

Tillé, Y. and M. Wilhelm (2017). Probability sampling designs: principles for choice of design and balancing. *Statistical Science 32*(2), 176–189.

Modélisation des profils respiratoires de patients sous oxygénothérapie

Juliana ALVES PEGORARO[1,2,3], Sophie LAVAULT [2,4], Nicolas WATTIEZ [2], Thomas SIMILOWSKI [2,4], Jésus GONZALEZ-BERMEJO [2,4] & Etienne BIRMELE [1,5]

[1] *UMR CNRS 8145, Laboratoire MAP5, Université de Paris, 45 rue des Saints-Pères, 75006, Paris, juliana.alves-pegoraro1@u-paris.fr, etienne.birmele@parisdescartes.fr*
[2] *Sorbonne Université, INSERM, UMRS1158 Neurophysiologie Respiratoire Expérimentale et Clinique, F-75005, Paris, nicolas.wattiez@upmc.fr, thomas.similowski@upmc.fr,*
[3] *SRETT, 11 Rue Heinrich, 92100, Boulogne-Billancourt*
[4] *AP-HP, Groupe Hospitalier Universitaire APHP-Sorbonne Université, site Pitié-Salpêtrière, Service de Pneumologie, Médecine Intensive et Réanimation (Département R3S), F-75013, sophie.lavault@aphp.fr, jesus.gonzalez@aphp.fr*
[5] *Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS, 7 rue René-Descartes, 67000, Strasbourg*

Résumé. La Bronchopneumopathie chronique obstructive (BPCO) est un important enjeu de santé publique. Cette maladie chronique est parmi les principales causes de mortalité dans le monde. Les patients atteints de la BPCO présentent une limitation importante de la fonction respiratoire avec des conséquences majeures sur la qualité de vie. Les exacerbations sont des évènements aigus caractérises par une aggravation des symptômes au delà des variations quotidiennes. A chaque nouvelle exacerbation, la qualité de vie se dégrade et les chances d'une nouvelle exacerbation augmentent. Certains auteurs ont montré une augmentation de la fréquence respiratoire moyenne dans les jours précédant une exacerbation, ouvrant la possibilité de l'identification précoce de ces évènements. Néanmoins, les résultats obtenus présentent encore des performances faibles. Dans ce contexte, nous nous sommes intéressés sur comment enrichir cet apprentissage grâce à l'utilisation d'un dispositif de suivi en continu. Dans un premier moment, nous étudions la possibilité de prendre en compte des variables autres que la fréquence respiratoire à fin de mieux décrire les signaux respiratoires avec des approches supervisées et non supervisées. Il ressort de cette première partie que l'amplitude inspiratoire permet d'améliorer la classification des séquences de respirations tout en consommant peu de ressources en mémoire et capacité de calcul. Une fois les variables à surveiller (débit d'oxygène, fréquence respiratoire et amplitude de l'inspiration mesurées toutes les cinq minutes) choisies et implémentées dans le dispositif, nous démontrons qu'il est possible de décrire des profils respiratoires individuels à travers de modèles de Markov à états cachés, qui permettent de mettre en évidence l'individualité des patients, tout en tenant en compte deux état physiologique distincts : le repos et l'effort.

Mots-clés. Télésurveillance, détection de nouveauté, modèle de Markov caché, profil respiratoire

Abstract. Chronic Obstructive Pulmonary Disease (COPD) is an important public health issue. This chronic disease is among the leading causes of death in the world. Patients with COPD present a significant limitation of respiratory function with major consequences on the quality of life. Exacerbations are acute events characterized by worsening of symptoms beyond daily variations. With each new exacerbation, the quality of life deteriorates and the chances of a new exacerbation increase. Some authors have shown an increase in the mean respiratory rate in the days preceding an exacerbation, opening the possibility of early identification of these events. Nevertheless, the results obtained still show poor performance. In this context, we are interested in how to enrich this learning through the use of a continuous monitoring system. First, we study the possibility of taking into account variables other than the respiratory rate in order to better describe the respiratory signals in supervised and unsupervised approaches. It emerges from this first part that the inspiratory amplitude makes it possible to improve the classification of the sequences of breaths while consuming few resources in memory and computing capacity. Once the variables to be monitored chosen and implemented in the device (oxygen flow, respiratory rate and amplitude of inspiration measured every five minutes), we demonstrate that it is possible to describe individual respiratory patterns through hidden Markov models, which make it possible to highlight the individuality of the patients, while taking into account two distinct physiological states : rest and effort.

Keywords. Telemonitoring, novelty detection, hidden Markov model, respiratory pattern

1 Introduction

La Bronchopneumopathie chronique obstructive (BPCO) est une maladie respiratoire chronique évitable et traitable, caractérisée par une obstruction permanente et progressive des voies aériennes. Dans le monde, la BPCO est parmi les principales causes de morbidité et de mortalité, représentant un problème de santé publique majeur [Global Initiative for Chronic Obstructive Lung Disease 2017 ; Rabe et al. 2007].

La BPCO résulte en un déclin accéléré de la fonction respiratoire chez un grand nombre de malades [Rabe et al. 2007]. L'évolution de la difficulté respiratoire mène souvent à une réduction des activités quotidiennes et à la dégradation de la qualité de vie.

Les périodes de détérioration aiguë des symptômes sont appelées exacerbations. Les exacerbations sont définies par une aggravation de l'état d'un patient au-delà des variations quotidiennes, d'une durée supérieure à 48 heures ou conduisant à une modification du traitement habituel [Société de Pneumologie de Langue Française 2010]. Les exacerbations mènent à une aggravation importante de la maladie. A chaque nouvelle

exacerbation, les chances de faire une nouvelle exacerbation et le risque de mortalité augmentent, au même temps que les fonctions pulmonaires se débilitent. L'identification précoce d'une exacerbation permet de débuter le traitement plus tôt, ce qui a une influence significative pour une récupération rapide [Wilkinson et al. 2004].

Plus précisément, certains auteurs ont trouvé une corrélation entre l'augmentation de la fréquence respiratoire et un événement d'exacerbation [Yañez et al. 2012 ; Borel et al. 2015]. Si les spécificités sont bonnes (93% et 89,7% respectivement), les sensibilités sont cependant faibles (respectivement 66% et 46,2%).

Par ailleurs, dans l'optique d'une médecine personnalisée et la moins invasive possible, il est souhaitable de recueillir des données à domicile et sans intervention nécessaire ni du patient ni du corps médical. Ceci est possible notamment via le dispositif TeleOx®, habituellement utilisé pour le télésuivi de patients sous oxygénothérapie de longue durée et qui enregistre des signaux de débit et pression dans le circuit à oxygène pendant $45s$ toutes les cinq minutes. Ce signal est une version très bruitée de la respiration, mais obtenue tout au long de l'utilisation de l'oxygène et sans manipulation. Des versions adaptées du firmware ont été utilisées afin soit de stocker les signaux bruts de pression et débit à 10 Hz, soit d'en extraire les fréquences et amplitudes de respiration médianes. Tous les protocoles ont été approuvés par des comités compétents et tous les participants ont fourni un consentement éclairé de participation à l'étude.

La présente étude se propose d'explorer les signaux respiratoires recueillis par ce dispositif afin de 1) déterminer si la prise en compte de variables autres que la fréquence respiratoire permettent d'améliorer la classification des séquences de respirations et 2) modéliser les profils respiratoires individuels des patients.

2 Choix des variables

Une question qui se pose en amont de la prédiction d'exacerbations est de définir comment résumer chaque portion de $45s$ de signal afin de bien caractériser un changement de respiration tout en étant le plus économe possible, TeleOx® étant limité en mémoire et capacité de calcul.

En raison de la rareté des évènements d'exacerbations, l'étude sur le choix de variables a été mené sur un autre type de changement de l'équilibre charge-capacité des muscles respiratoires : l'effort physique. Les signaux bruts de 20 sujets sains et de 8 patients atteints de la BPCO ont été enregistrés sous deux protocoles permettant d'isoler, de façon certaine pour les sujets sains et de façon approximative pour les sujets BPCO, des périodes de repos et des périodes d'effort.

Chaque fenêtre de $45s$ de mesure, pour laquelle un label *repos* ou *effort* est disponible, est résumée par quatre valeurs possibles :

— la fréquence respiratoire, calculée comme l'inverse de la médiane des longueurs des respirations identifiées [Soler et al. 2019]

— l'amplitude, correspondant à la médiane des amplitudes à l'inspiration, calculées comme la distance entre les minima du signal de pression et une ligne de base estimée
— un vecteur $(\hat{\mu}, \hat{\Phi}, \hat{\theta}) \in R^3$ correspondant à l'estimation d'un modèle ARIMA(1,1,1).
— le vecteur de sa transformation de Fourier discrète, limité aux fréquences inférieures à 2Hz.

Afin d'identifier les indicateurs les plus efficaces pour différencier repos et effort, tout en limitant l'influence de la méthode de classification choisie, deux approches sont étudiées.

— une classification supervisée est menée à l'aide de SuperLearner [van der Laan, E. C. Polley et Hubbard 2007 ; E. Polley et al. 2018]. Cette méthode nous permet de combiner un ensemble de modèles, dont différents modèles linéaires généralisés, des k-NN à noyau, des forêts aléatoires ou XGBoost, via une combinaison linéaire dont les poids sont estimés par validation croisée.
— une approche *one-class* plus proche du cas d'utilisation pratique, où seules des données au repos sont utilisées pour l'apprentissage. La méthode retenue est alors basée sur la distance de Mahalanobis [McLachlan 1999 ; Aggarwal 2017].

Pour les sujets sains, la fréquence respiratoire seule obtient la pire performance pour la classification des périodes de repos et effort (AUC respectives de 0.69 et 0.68 en supervisé ou one-class), alors que tous les autres indicateurs, seuls ou combinés, présentent des meilleurs résultats. Les coefficients de Fourier ou la combinaison de la fréquence respiratoire avec l'amplitude obtiennent par exemple de bien meilleures performances (AUC au-delà de 0.98 ou de 0.90).

Les performances sur les données des patients atteints de la BPCO sont inférieures à celles obtenues sur les données des sujets sains. Cela peut être expliqué par plusieurs facteurs qui demanderont à être confirmés : i. les horaires concernant les périodes repos et sport sont approximatifs ii. pour certains patients, grands insuffisants respiratoires, toute activité du quotidien, comme se lever, prendre une douche ou marcher, peut être déjà considérée comme leur effort physique maximal iii. à l'opposé, l'activité physique programmée pendant la journée est optimisée par un professionnel de santé pour ne pas dépasser leurs limitations respiratoire et musculaire, et donc paradoxalement, peut être un moindre effort que les mouvements du quotidien.

Malgré ces limites, on retrouve aussi que la performance obtenue avec la fréquence respiratoire comme seul indicateur est à nouveau très inférieure à celles obtenues en la combinant avec l'amplitude et/ou les coefficients ARIMA, ou en considérant les coefficients de Fourier.

En première conclusion, il ressort de cette étude que la fréquence respiratoire seule est un indicateur insuffisant pour la détection automatique de changements de la respiration. L'ajout de n'importe quel indicateur décrit ci-dessus est capable d'améliorer le pouvoir de classification, notamment celui de l'amplitude qui est implémentable dans TeleOx® sans en modifier les caractéristiques techniques pour avoir la possibilité de modéliser des profils respiratoires individuels.

3 Profils respiratoires individuels

Basé sur les conclusions précédentes, des nouveaux enregistrements ont été réalisés en utilisant la nouvelle version du firmware de TeleOx®, capable d'enregistrer des mesures de débit d'oxygène, fréquence respiratoire et amplitude de l'inspiration toutes les cinq minutes. Pour cette partie de l'étude, les patients ont été équipés avec deux dispositifs TeleOx® : un pour une source fixe, utilisée exclusivement dans la chambre du patient, et un pour une source portable, utilisée lors des déplacements. Les données acquises sur les deux TeleOx® mais correspondant à un même patient sont combinées et triées par date.

La base de données étudiée comprend 27 enregistrements. Les durées de suivi étant très variables, seulement les 14 derniers jours d'enregistrements de chaque patient sont gardés pour les analyses suivantes.

Afin de prendre en compte le débit d'oxygène utilisé, variable au long du temps et influent sur le profil respiratoire des patients, on propose un nouvel indicateur, appelé l'oxygénation estimée, défini par $d\acute{e}bit * \sqrt{amplitude}$.

Ensuite, les séries temporelles de fréquence et oxygénation estimée acquises pendant les 7 premières journées sont utilisées pour entraîner des modèles de Markov à 2 états cachés [Rabiner et Juang 1986] individuels. On émet l'hypothèse que, si ces deux états cachés correspondent aux états repos et effort, les paramètres du modèle ont des sens réels :

- Les probabilités de transition correspondent aux probabilités de passer d'un état à l'autre. Il est attendu que les probabilités de rester à l'effort alors qu'on est à l'effort soit supérieure à la probabilité de changer d'état. Idem pour le repos.
- Les distributions d'émission donnent la relation entre les états cachés et les observations. Il est espéré que les observations en fréquence respiratoire et oxygénation estimée soient différentes selon l'état physiologique.

Les séries appartenant à la deuxième semaine ont les états cachés prédits selon ce modèle individuel. Ces prédictions sont comparées aux sources à oxygène étant à l'origine de chaque mesure : la source portable correspond à l'effort et la fixe au repos. Encore une fois, ces labels sont approximatifs, le patient pouvant réaliser des activités physiques à l'intérieur de sa chambre ou bien se reposer à l'extérieur.

L'AUC obtenue pour tous les patients confondus est de 0,72, indiquant que cette modélisation permet de décrire le profil respiratoire d'un individu par rapport à ses états physiologiques de repos et effort. Il ressort également l'importance de la modélisation personnalisée, les paramètres estimés étant différents d'un patient à l'autre.

4 Perspectives

Ces études préliminaires permettent de proposer un outil et un algorithme pour mettre en place un suivi à long terme. Pour optimiser le télésuivi, modéliser l'état de base d'un

patient permettrait de quantifier l'évolution journalière du profil respiratoire et, potentiellement, l'identification d'un changement important de l'état de santé, comme c'est le cas de l'exacerbation. La modélisation de l'état stable n'a pas été possible dans cette étude puisque l'inclusion des patients a été réalisée alors qu'ils étaient tous en post-exacerbation, mais d'autres travaux sont en cours pour confirmer cette hypothèse.

Bibliographie

Aggarwal, C. C. (2017). *Outlier Analysis*. 2nd. Springer Publishing Company, Incorporated.

Borel, J. C. et al. (2015). "Parameters recorded by software of non-invasive ventilators predict COPD exacerbation: a proof-of-concept study". *Thorax* 70.3, p. 284-286.

Global Initiative for Chronic Obstructive Lung Disease (2017). *Pocket guide to COPD diagnosis, management, and prevention.*

McLachlan, G. J. (1999). "Mahalanobis distance". *Resonance* 4 (6), p. 20-26.

Polley, E. et al. (2018). *SuperLearner: Super Learner Prediction*. R package version 2.0-24. url : `https://CRAN.R-project.org/package=SuperLearner`.

Rabe, K. F. et al. (2007). "Global Strategy for the Diagnosis , Management , and Prevention of Chronic Obstructive Pulmonary Disease GOLD Executive Summary". 176, p. 532-555.

Rabiner, L. et Juang, B. (1986). "An introduction to hidden Markov models". *IEEE ASSP Magazine* 3.1, p. 4-16.

Société de Pneumologie de Langue Française (2010). "Recommandation pour la Pratique Clinique : Prise en charge de la BPCO". *Revue des Maladies Respiratoires* 27, p. 522-548.

Soler, J. et al. (2019). "Validation of respiratory rate measurements from remote monitoring device in COPD patients". *Respiratory Medicine and Research* 76, p. 1-3.

van der Laan, M. J., Polley, E. C. et Hubbard, A. E. (2007). "Super Learner". *Statistical Applications in Genetics and Molecular Biology* 6.1, Article25.

Wilkinson, T. M. A. et al. (2004). "Early Therapy Improves Outcomes of Exacerbations of Chronic Obstructive Pulmonary Disease". 169, p. 1298-1303.

Yañez, A. M. et al. (2012). "Monitoring Breathing Rate at Home Allows Early Identification of COPD Exacerbations". *Chest* 142.6, p. 1524-1529.

Bayesian block-diagonal graphical models via the Fiedler prior

Julyan Arbel[1] & Mario Beraha[2,3] & Daria Bystrova[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France*
{julyan.arbel, daria.bystrova}@inria.fr
[2] *Department of Mathematics, Politecnico di Milano*, mario.beraha@polimi.it
[3] *Department of Computer Science, Universitá degli studi di Bologna*

Résumé. Nous étudions le problème de l'inférence de la structure d'indépendance conditionnelle entre les entrées d'un vecteur aléatoire gaussien, principalement dans le but d'obtenir des groupes de variables indépendantes. Cela peut se traduire par l'estimation d'une matrice de précision (inverse de la matrice de covariance) avec une structure bloc-diagonale. Cette approche se base sur des techniques de théorie spectrale des graphes et de clustering spectral. Nous proposons une nouvelle loi a priori, le prior de *Fiedler*, qui satisfait une propriété de *shrinkage* vers les matrices de précision à structure bloc-diagonale. Nous comparons le *shrinkage* induit par ce prior de Fiedler et par le Graphical Lasso, et comparons leurs performances sur un ensemble de données simulées.

Mots-clés. Modèles graphiques, matrice de précision, valeur de Fiedler, théorie spectrale des graphes.

Abstract. We study the problem of inferring the conditional independence structure between the entries of a Gaussian random vector. Our focus is on finding groups of independent variables. This can be translated into the estimation of a precision matrix (inverse of the covariance matrix) with a block-diagonal structure. We borrow ideas from spectral graph theory and spectral clustering and propose a novel prior called *Fiedler* prior showing shrinkage properties towards block-diagonal precision matrices. We compare the shrinkage induced by our prior and the popular Graphical Lasso prior, and compare their performance on a simulated dataset.

Keywords. Graphical models, Precision matrix, Fiedler value, Spectral graph theory.

1 Introduction

Understanding the dependence structure among large numbers of variables is an important topic in many different application areas, such as ecology, neuroscience, genetics. In a graphical model, the dependence structure of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ can be represented by a graph G with nodes $\{1, \ldots, p\}$, where each node i corresponds to a random variable Y_i and edges represent the probabilistic relationships between nodes.

If there is not an edge connecting nodes i and j it means that, conditionally on $\boldsymbol{Y} \setminus \{Y_i, Y_j\}$, Y_i and Y_j are independent. When $\boldsymbol{Y}$ is assumed to be a Gaussian random vector, the objective of the inference is the precision matrix Σ^{-1}, the inverse of the covariance matrix, which encodes conditional (in)dependencies: $\boldsymbol{\Sigma}^{-1}_{ij} = 0$ if and only if Y_i and Y_j are independent given $\boldsymbol{Y} \setminus \{Y_i, Y_j\}$. For a recent review, see Maathuis et al. (2018).

In the Bayesian setting, a prior distribution is assumed on Σ^{-1} that encourages its off-diagonal entries to be zero or close to zero. Two strategies are commonly employed: shrinkage priors and graph-based priors. The former approach can be understood as a generalization of commonly used shrinkage priors (such as the Lasso prior) in linear regression to positive definite matrices. See, for instance, Wang (2012); Li et al. (2019) and references therein. In the latter approach, instead, a prior is assumed for G and, conditionally to G a prior on Σ^{-1} is assumed such that an absence of the edge between nodes i and j in G implies $\Sigma^{-1}_{ij} = 0$. See Mohammadi and Wit (2015) and references therein. Each approach has its pros and cons. Generally speaking, posterior inference in graph-based models is less efficient because they require transdimensional Markov chain Monte Carlo (MCMC) sampling strategies (Green, 1995) in a huge dimensional parameter space. On the other hand, models based on shrinkage priors usually lead to simpler and more efficient MCMC algorithms, but the estimates of G obtained from Σ^{-1} might be worse (Mohammadi and Wit, 2015).

We propose a novel shrinkage prior for Bayesian graphical modeling, called Fiedler prior, which is particularly useful for estimating sparse precision matrices Σ^{-1} with a block-diagonal structure. We borrow ideas from spectral clustering (Von Luxburg, 2007) and define the prior based on the spectrum of a transformation of the precision matrix. This allows Fiedler prior to enforce block-diagonal structure on the precision matrix.

There exist several methods for sparse covariance matrix estimation based on approximating the precision matrix in a block-diagonal way. These approaches usually follow a two-step procedure, first detecting the block-diagonal structure and then applying the Graphical Lasso (hereafter referred to as G-Lasso) algorithm to each block for estimating the precision matrix (see eg Devijver and Gallopin, 2018).

2 The Bayesian graphical model

Before presenting the main contribution of this work, let us introduce some preliminary definitions and results.

2.1 Graph Laplacian and Fiedler value

Given a weighted graph with weights $W = \{w_{ij}\}_{i,j=1}^{p}$, $w_{ij} \geq 0$, define its unnormalized Laplacian as $L = D - W$, where $D = \text{diag}(\sum_j w_{1j}, \ldots, \sum_j w_{pj})$. The analysis of the eigenvalues $\lambda_1 \leq \ldots \leq \lambda_p$ of L and the associated eigenvectors is formalized in the field of *spectral graph theory*, cf. Spielman (2012).

It is well known that $\lambda_1 = 0$ for any L. The multiplicity of the eigenvalue 0 corresponds to the number of connected components in the graph (see, e.g., Von Luxburg, 2007, Proposition 2). In particular, the graph is connected if and only if the second smallest eigenvalue of L, known as the Fiedler value or algebraic connectivity, satisfies $\lambda_2 > 0$. Moreover, the eigenspace associated with 0 is spanned by the indicator vectors of those components. This is the key motivation underlying spectral clustering.

2.2 The Fiedler prior

In this section, we formalize the Fiedler prior, a prior over partial correlation matrices. The entries of the partial correlation matrix Ω are elements of $[-1, 1]$ which are expressed as a function of Σ^{-1} by

$$\omega_{ij} = -\Sigma^{-1}_{ij} \Big/ \sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}. \tag{1}$$

Formally, let $L(|\Omega|)$ be the Laplacian matrix associated to the matrix $|\Omega|$ with entries $|\omega_{ij}|$, and let $\lambda_1(\Omega), \ldots, \lambda_p(\Omega)$ denote its eigenvalues. Then Ω follows the Fiedler prior with parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ if it has density

$$p(\Omega|\boldsymbol{\delta}) = \frac{1}{Z} \exp\Big(- \sum_{j=1}^{p} \delta_j \lambda_j(\Omega)\Big) \tag{2}$$

with respect to the standard Lebesgue measure on the space of $[-1, 1]$-valued symmetric matrices. Note that $Z = Z(\boldsymbol{\delta})$ is finite almost surely because the support of $p(\Omega)$ is bounded. Since $\lambda_1 = 0$ for any Ω, we will always set $\delta_1 = 0$. The original idea that initiated the definition and study of this prior is the use of the Fiedler value for penalized maximum likelihood estimation in neural networks (Tam and Dunson, 2020).

To transpose (2) to precision matrices, we use an approach similar to the one in Barnard et al. (2000), who instead work on the covariance matrix Σ. We decompose Σ^{-1} into a partial correlation matrix and an inverse-scale matrix: $\Sigma^{-1} = T\Omega T$, where $T = \mathrm{diag}(\tau_1, \ldots, \tau_p)$. The conditional dependencies can be read equivalently from Ω or Σ^{-1}. Our prior specification is completed by assuming

$$\tau_j \overset{\text{iid}}{\sim} \mathrm{Exp}(\eta), \quad j = 1, \ldots, p, \tag{3}$$

where $\mathrm{Exp}(\eta)$ is the exponential distribution with mean η^{-1}.

As a simple illustration, we compare the marginal distribution on the off-diagonal entries ω_{ij} under G-Lasso and Fiedler priors. Since both priors involve intractable normalizing constants, we use an MCMC algorithm to sample from them. In particular, for the G-Lasso prior we simulate from the prior on Σ^{-1} defined in Wang (2012) and compute ω_{ij} as in (1). For both priors we assume $p = 15$, for the Fiedler prior we fix $\boldsymbol{\delta} = (0, \delta, \delta, \delta, 0, \ldots, 0)$ with $\delta = 25$. For the G-Lasso prior, we employ a double exponential kernel with parameter λ. Figure 1 shows the marginal distributions for different

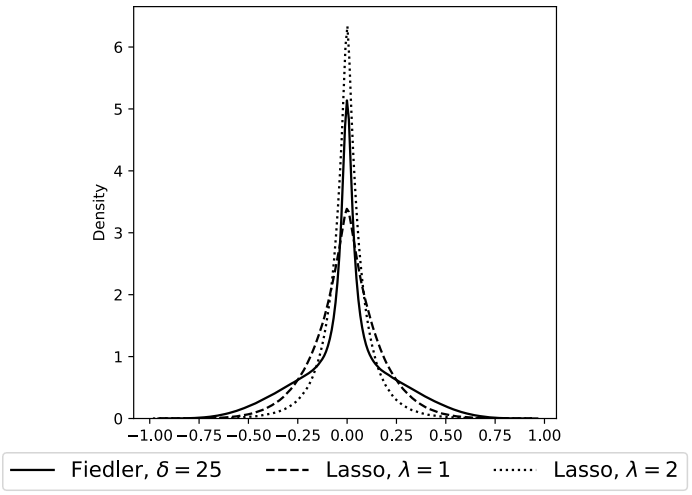


Figure 1: Marginal priors for the off-diagional entries $\{\omega_{ij}, j > i\}$ under the Fiedler prior and the G-Lasso prior, for different values of the hyperparameters.

values of the parameters. Note that the G-Lasso prior shows the usual tradeoff between local and global shrinkage: to obtain shrinkage for values that are close to 0, also the values that are far from 0 are significantly shrunk (see the tails for $\lambda = 2$). On the contrary, observe how the tails of the Fiedler prior are significantly heavier than the ones of the G-Lasso for both choices of λ, showing that good shrinkage of small values can be achieved without overshrinking the signal of large values.

3 Numerical illustrations

We present a simple simulation study to show the difference between the Fiedler prior and the G-Lasso. We simulated $n = 250$ observations independently from a six-dimensional zero centered normal distribution with precision matrix equal to

$$\Sigma^{-1} = \begin{bmatrix} A, & \mathbf{0} \\ \mathbf{0}, & A \end{bmatrix}, \qquad A = \begin{bmatrix} 3, & 1.5, & 1.5 \\ 1.5, & 3, & 1.5 \\ 1.5, & 1.5, & 3 \end{bmatrix}. \tag{4}$$

Such a model separates the variables into two blocks: the first three and the last three.

We considered different prior specifications for Σ^{-1}. A "well-specified" and a "misspecified" Fiedler prior with respective parameters $\boldsymbol{\delta} = (0, \delta, 0, \ldots, 0)$ and $\boldsymbol{\delta} = (0, \delta, \delta, \ldots, 0)$. Such parameters $\boldsymbol{\delta}$ imply that they make use of only λ_2, or both λ_2 and λ_3, respectively, which should yield two or three separate groups of variables, respectively. Finally, we considered the G-Lasso with double exponential parameter λ.

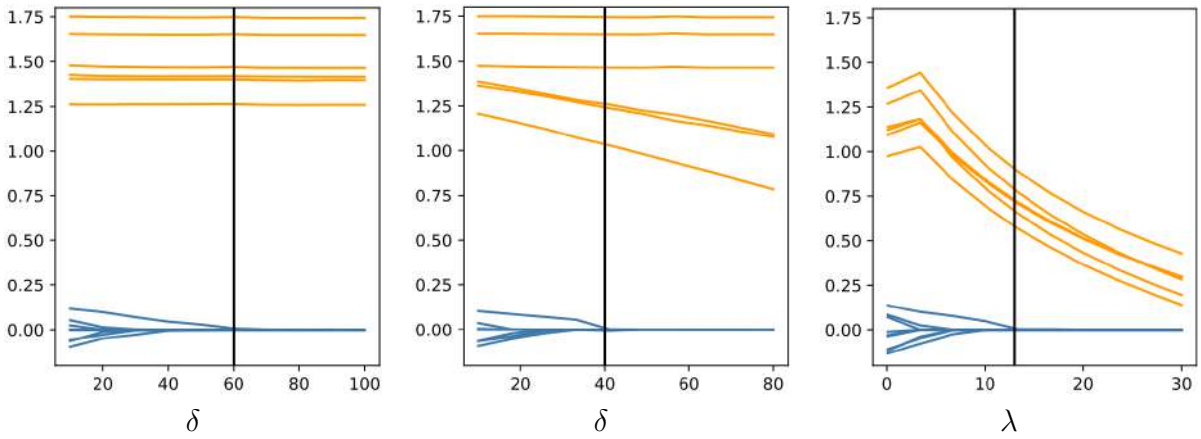


Figure 2: Coefficients of the MAP estimates of Σ^{-1} as a function of the values δ or λ under the tree models: from left to right, well-specified Fiedler, misspecified Fiedler, and G-Lasso. Orange lines and blue lines refer to the estimates of the nonzero and zero off-diagonal elements of Σ^{-1}, respectively. The vertical black line indicates when all the estimates of the zero entries in (4) are below 10^{-5} in absolute value.

We computed the maximum a posteriori (MAP) estimate of Σ^{-1} for various values of δ and λ and looked at the values of the entries of Σ^{-1} as a function of δ and λ.

Figure 2 reports the plots of the "paths" for the tree priors employed. The G-Lasso prior shows the usual overshrinking phenomenon: to estimate values close to zero for the zeros in Σ^{-1}, all the values are shrunk to small values. The well-specified Fiedler behaves correctly: it shrinks to zero the correct terms in Σ^{-1} without "penalizing" the nonzero entries. This shows exactly how the Fiedler prior works: it encourages sparsity only to separate components of the graph associated with Σ^{-1}. Once the components are separated, the other variables are free to assume any large value. Finally, the misspecified Fiedler shows an in-between behavior. In order to recover the two-block structure in Σ^{-1} also some of its nonzero entries are shrunk. This suggests that great care must be taken in carefully choosing the parameter $\boldsymbol{\delta}$.

4 Discussion and future work

In this work, we presented a novel prior for partial correlation matrices, namely the Fiedler prior, and showed an application to Bayesian graphical modeling for Gaussian variables. The Fiedler prior is particularly suited to detect block-diagonal structures.

Several interesting questions are still open. First of all, the choice of parameter $\boldsymbol{\delta}$ seems to be crucial. Assuming a prior distribution on it is unpractical due to the intractable normalizing constant in (2). Hence, a suitable prior elicitation strategy, as well as sensitivity analysis, must be devised. Second, MCMC computation based on gradient

information is burdensome due to the need of computing the gradient of the eigendecomposition of Ω. To this end, we might exploit an approximate gradient formulation based on the Rayleigh quotient characterization, as done in Tam and Dunson (2020).

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 1281–1311.

Devijver, E. and Gallopin, M. (2018). "Block-diagonal covariance selection for high-dimensional Gaussian graphical models." *Journal of the American Statistical Association*, 113(521), 306–314.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4), 711–732.

Li, Y., Craig, B. A., and Bhadra, A. (2019). "The graphical horseshoe estimator for inverse covariance matrices." *Journal of Computational and Graphical Statistics*, 28(3), 747–757.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Mohammadi, A. and Wit, E. C. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1), 109–138.

Spielman, D. (2012). "Spectral graph theory." *Combinatorial scientific computing*, (18).

Tam, E. and Dunson, D. (2020). "Fiedler Regularization: Learning Neural Networks with Graph Sparsity." In *Proceedings of ICML 37*, volume 119, 9346–9355. PMLR.

Von Luxburg, U. (2007). "A tutorial on spectral clustering." *Statistics and Computing*, 17(4), 395–416.

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computation." *Bayesian Analysis*, 7(4), 867–886.

ESTIMATION OF THE COVARIATE CONDITIONNAL TAIL EXPECTATION : A DEPTH-BASED LEVEL SET APPROACH

Elisabeth Armaut[1], Roland Diel[2] & Thomas Laloë[3]

[1] *Université de Nice Côte d'Azur, armaut@unice.fr*
[2] *Université de Nice Côte d'Azur, Roland.Diel@unice.fr*
[3] *Université de Nice Côte d'Azur, Thomas.Laloe@unice.fr*

Résumé. De nos jours, dans pratiquement tous les domaines tels que la finance, la médecine, l'écologie, l'industrie..., il est impossible d'éviter des risques ! Par exemple, en finance, le "risque" signifie souvent la possibilité de perdre de l'argent. En hydrologie, le risque peut par ailleurs représenter la quantité d'eau dépassant le niveau de remplissage maximum d'un barrage. La *Conditionnal Covariate Tail Expectation* (ou CCTE) est une mesure de risque qui quantifie un coût moyen associé à $d \geq 1$ facteurs de risque non nécessairement homogènes. Dans notre cadre d'étude, la zone de risque est représentée par un ensemble de niveau inférieur associé à une fonction de profondeur statistique multivariée. Nous proposons un estimateur consistant de la CCTE avec une vitesse de convergence : cet estimateur fait intervenir une estimation de l'ensemble de niveau associé à la profondeur en question via une méthode *plug-in*. Une étude sur simulation vient compléter l'étude des performances de notre estimateur.

Mots-clés. Estimation *plug-in*, fonction de profondeur multivariée, théorie du risque.

Abstract. Nowadays, in almost all fields such as finance, medicine, ecology, industry.., it is not possible to avoid risks. For instance, in finance, risk often means that there is potential for money loss. In hydrology, risk could represent the amount of water which exceeds the maximum storage level of a dam. The *Conditionnal Covariate Tail Expectation* (CCTE) is a risk measure that quantifies an expected cost associated to $d \geq 1$ risk factors which are heterogeneous in nature. In our setting, the risk region in the problem at hand is represented by a depth-based lower level set. We provide a consistent estimator of the CCTE with a rate of convergence : this estimator involves a *plug-in* approach when estimating the lower level set. A simulation study complements the performances of our estimator.
Key-Words. *Plug-in* estimation, multivariate depth function, risk theory.

1 Introduction

La théorie du risque est une branche de la statistique qui s'intéresse aux évènements peu probables. Le but principal recherché est de pouvoir gérer la part d'incertude de certains évènements, ce qui permet la mise en place de mesures de prévention. La théorie du risque est appliquée dans divers domaines, par exemple, en hydrologie pour prévoir les crues, en finance pour protéger la valeur du portefeuille après un investissement, etc...

Usuellement, une mesure de risque est une application définie sur un ensemble de variables aléatoires à valeurs réelles. Il est à noter que, pour un phénomène, la prise en compte d'un seul facteur de risque pourrait considérablement affectée l'exactitude du modèle à l'étude. Modéliser la structure de dépendance de données multivariées permet alors d'obtenir des résultats significatifs et réprésentatifs dans l'analyse du risque. Par exemple en hydrologie, plusieurs phénomènes sont décrits par le biais de deux ou plusieurs variables corrélées. Ces dernières sont considérées jointement afin de représenter efficacement ces phénomènes hydrologiques. Ainsi, la probabilité de réalisation d'un risque ne peut pas être estimée sur la base de l'analyse univariée. La littérature de risques hydrologiques a été largement étudiée dans un cadre multivarié et traite principalement un ou plusieurs des éléments suivants : (1) montrer l'importance et l'utilité du cadre multivarié, (2) trouver un modèle de distributions multivariées approprié afin de modéliser les risques, et (3) définir et étudier des temps de retour miltivariés (cf. Chebana and Ouarda (2011)). L'étude de risques amène à l'étude de "régions" quantiles : celle des risques univariés via des quantiles univariés a été largement traitée dans la littérature. Quant aux risques multivariés, l'étude de quantiles multivariés a gagné beaucoup d'attention ces dernières décénies, notamment les quantiles basés sur une loi de probabilité multivariée (cf. Belzunce et al. (2007), Dehaan and Huang (1995), Cousin and Di Bernardino (2013)), ou encore les quantiles basés sur une fonction de profondeur (cf. Zuo and Serfling (2000)).

Il est pertinent d'analyser le comportement d'un coût modélisé par une variable aléatoire réelle Y (par exemple, la somme d'argent gagnée ou perdue dans un investissement sur une certaine période), et ce par rapport à $d \geq 1$ facteurs de risque différents $\boldsymbol{X} \in \mathbb{R}^d$. Intuitivement, la variable de coût Y va dépendre des facteurs de risques du phénomène étudié. Dans un cadre d'étude général, Laloë et al. (2015) proposent d'analyser le comportement d'une variable aléatoire réelle Y qui dépend d'un vecteur aléatoire de risques $\boldsymbol{X} \in \mathbb{R}^d$. Plus précisément, la *Covariate Conditionnal Tail Expectation* suivante définie une mesure de risque :

$$\mathrm{CCTE}_{F,\alpha}(Y, \boldsymbol{X}) := \mathbb{E}[Y | \boldsymbol{X} \in \mathcal{L}_{F_{\boldsymbol{X}}}(\alpha)], \alpha \in (0,1), \tag{1.1}$$

où,

$$\mathcal{L}_{F_{\boldsymbol{X}}}(\alpha) := \left\{ x \in \mathbb{R}^d : F_{\boldsymbol{X}}(x) \geq \alpha \right\}, \tag{1.2}$$

est l'ensemble de niveau (supérieur) associé à la fonction de répartion $F_{\boldsymbol{X}} := F$ du v.a $\boldsymbol{X} \in \mathbb{R}^d$.

2 *Depth-based Covariate Conditionnal Tail Expectation*

L'utilisation d'une mesure de risque basée sur les ensembles de niveau d'une fonction de répartition nous restreint à considérer uniquement des orientations particulières du risque (cf. Figure 2, graphe gauche). Par exemple, on peut considérer des températures très basses ou très élévées, mais pas les deux à la fois. Afin de contourner ce problème de dépendance en une direction canonique, au lieu de considérer des ensembles de niveau de la forme $\mathcal{L}_{F_{\boldsymbol{X}}}(\alpha)$, des ensembles de niveau de la forme $\mathcal{L}_{F_{R\boldsymbol{X}}}(\alpha)$ ont été étudiés dans la littérature, où R est une matrice de rotation dans $\mathbb{R}^d$ (cf. Torres et al. (2020)). Une telle matrice

de rotation R permettrait de mettre en évidence d'autres zones de risque dans $\mathbb{R}^d$; cependant, dans ce cas, la CCTE dépendrait encore d'une orientation via la rotation R. Pour cela, afin d'étudier une mesure de risque complètement indépendante de l'orientation, nous proposons une approche nouvelle en considérant la CCTE pour des ensembles de niveau associés à une fonction de profondeur statistique multivariée (au sens de la Définiton 2.1 dans Zuo and Serfling (2000), cf. Figure 1), et qui est notée

$$D : \mathbb{R}^d \times \mathscr{P}(\mathbb{R}^d) \to \mathbb{R}_+,$$

avec $\mathscr{P} := \mathscr{P}(\mathbb{R}^d)$ désignant l'ensemble des mesures de probabilités sur $\mathbb{R}^d$. En efffet, une fonction de profondeur ordonne des données selon leur *degré de centralité*, et fournit un ordre statistique du *centre* vers l'extérieur (cf. Zuo and Serfling (2000)).

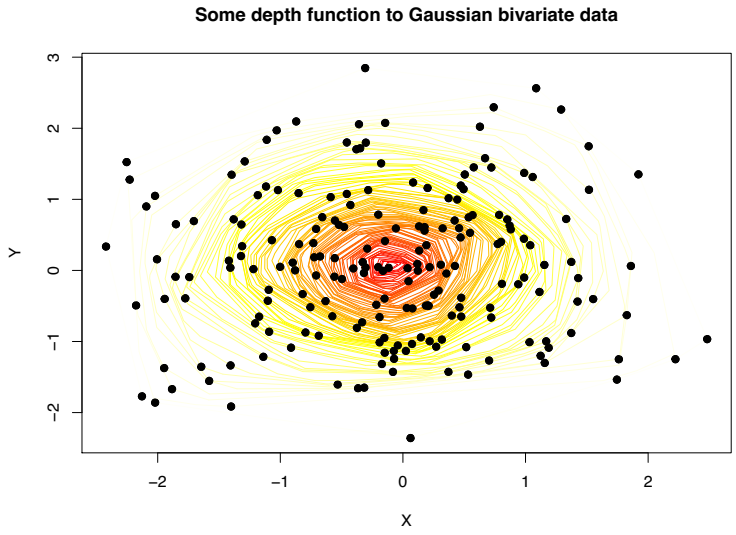


FIGURE 1 – Régions de profondeur faible à élévée (zones jaunes/blanches à rouges/oranges respectivement) pour des vecteurs gaussiens dans $\mathbb{R}^2$.

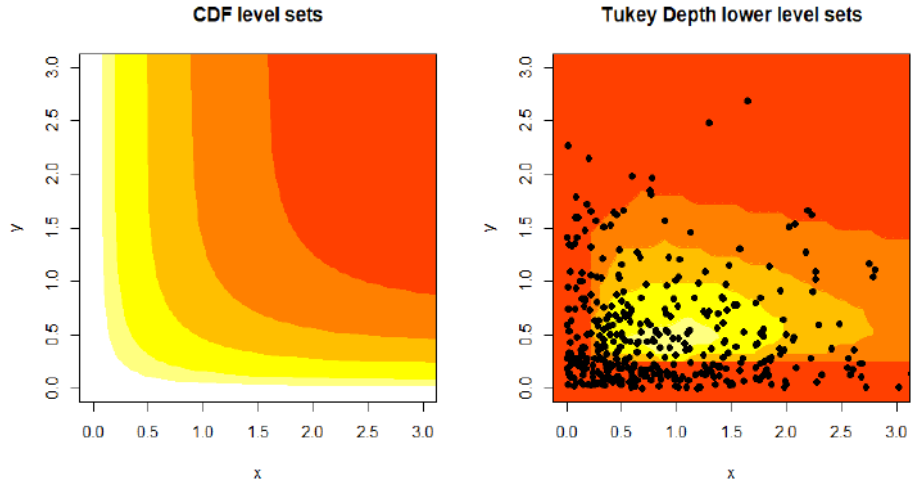


FIGURE 2 – Ensembles de niveaux (supérieurs) pour une fonction de répartition et ensembles de niveaux inférieurs pour une profondeur respectivement (de gauche à droite respectivement).

Afin d'étudier les zones de risque via une fonction de profondeur, il faut s'intéresser aux régions de faible profondeur sur lesquelles nous pouvont étudier le comportement de la covariable de coût Y. En d'autres termes, considérant un niveau $\alpha > 0$, une mesure de probabilité P sur $\mathbb{R}^d$, nous considérons l'ensemble de niveau

$$\boldsymbol{\mathcal{L}_D(\alpha) := \left\{ x \in \mathbb{R}^d : D(x, P) \leq \alpha \right\}}. \tag{2.1}$$

Étant donné une suite d'estimateurs consistants $(\tilde{P}_n)_{n\geq 1}$ de P, nous proposons d'estimer $\boldsymbol{\mathcal{L}_D(\alpha)}$ par l'ensemble

$$\boldsymbol{\mathcal{L}_{n,D}(\alpha) := \mathcal{L}_n(\alpha) = \left\{ x \in \mathbb{R}^d : D_n(x) := D(x, \tilde{P}_n) \leq \alpha \right\}, n \geq 1.} \tag{2.2}$$

3 Estimation et vitesse de convergence

Soit $\boldsymbol{X}$ un vecteur aléatoire de loi $P \in \mathscr{P}(\mathbb{R}^d)$, Y une va réelle, et $\alpha > 0$. Dans notre cadre d'étude, on s'interesse à l'estimation de la CCTE basée sur les ensembles de niveau d'une profondeur, définie par :

$$\mathrm{CCTE}_{D,\alpha}(Y, \boldsymbol{X}) := \mathbb{E}[Y | \boldsymbol{X} \in \mathcal{L}_D(\alpha)]. \tag{3.1}$$

Soient $n_1, n_2 \geq 1$. Soient deux échantillons,

$$\begin{aligned} \tilde{S}_{n_1} &:= (\tilde{\boldsymbol{X}}_i)_{i=1,..,n_1} \text{ de même loi que } \boldsymbol{X}, \text{ et} \\ S_{n_2} &:= ((Y_i, \boldsymbol{X}_i))_{i=1,..,n_2} \text{ de même loi que } (Y, \boldsymbol{X}), \end{aligned} \tag{3.2}$$

tels que les $(Y_i, \boldsymbol{X}_i)_i$ et les $(\tilde{\boldsymbol{X}}_i)_i$ soient indépendants. Considérons $\mathcal{L}_{n_1}(\alpha)$ un estimateur (calculé à partir de $\tilde{S}_{n_1}$) de $\mathcal{L}_D(\alpha)$. On peut alors construire à partir

de l'échantillon S_{n_2} un estimateur pour la CCTE (3.1) :

$$\widehat{\mathrm{CCTE}}_{D,\alpha}^{n_1,n_2}(Y,\boldsymbol{X}) := \mathbb{E}_{S_{n_2}}[Y|\boldsymbol{X} \in \mathcal{L}_{n_1}(\alpha)] = \frac{\sum_{i=1}^{n_2} Y_i \mathbb{1}_{\boldsymbol{X}_i \in \mathcal{L}_{n_1}(\alpha)}}{\sum_{i=1}^{n_2} \mathbb{1}_{\boldsymbol{X}_i \in \mathcal{L}_{n_1}(\alpha)}}, \tag{3.3}$$

sous réserve que, pour $n_1, n_2 \geq 1$, $\mathbb{P}$-p.s. $\sum_{i=1}^{n_2} \mathbb{1}_{\boldsymbol{X}_i \in \mathcal{L}_{n_1}(\alpha)} > 0$.

Notre premier résultat lie la vitesse de convergence de la $\widehat{\mathrm{CCTE}}$ à celle de $\mathcal{L}_n(\alpha)$. On suppose que :

(H0) il existe une suite de réels strictement positifs $(v_n)_{n\geq 1}$ telle que

$$v_n \cdot \mathbb{P}[\boldsymbol{X} \in \mathcal{L}_n(\alpha) \Delta \mathcal{L}_D(\alpha)] \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

où $A\Delta B = (A\backslash B) \cup (B\backslash A)$ désigne la différence symétrique entre les ensembles A et B.

On dérive alors la vitesse de convergence pour la $\widehat{\mathrm{CCTE}}$ dans le théorème suivant.

Théorème 1. *Soient $\alpha > 0$ et $P \in \mathscr{P}(\mathbb{R}^d)$. On suppose que*

$$\mathbb{P}[\boldsymbol{X} \in \mathcal{L}_D(\alpha)] > 0.$$

*Sous l'hypothèse **(H0)**, en supposant qu'il existe un $r \geq 2$ tel que Y est r-intégrable, on a :*

$$n^{\frac{r-1}{2r}} \wedge v_n^{\frac{r-1}{r}} \left| \widehat{\mathrm{CCTE}}_{D,\alpha}^{n}(Y,\boldsymbol{X}) - \mathrm{CCTE}_{D,\alpha}(Y,\boldsymbol{X}) \right| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Il est également possible de remplacer **(H0)** par une convergence au sens de la mesure de Lebesgue, à condition de faire des restrictions sur la loi de $\boldsymbol{X} \sim P$. On note λ_d la mesure de Lebesgue sur $\mathbb{R}^d$, et lorsque P est continue on note f la densité de $\boldsymbol{X}$:

(H1) : (i) $\exists$ $(v_n)_{n\geq 1}$ une suite de réels strictement positifs telle que

$$v_n \cdot \lambda_d(\mathcal{L}_n(\alpha) \Delta \mathcal{L}_D(\alpha)) \xrightarrow[n\to\infty]{\mathbb{P}} 0, \text{ et}$$

(ii) $\exists\, p > 1$, $\|f\|_{p,\lambda} := \left(\int_{\mathbb{R}^d} f(x)^p dx \right)^{\frac{1}{p}} < +\infty$.

Théorème 2. *Soit $\alpha > 0$, $P \in \mathscr{P}(\mathbb{R}^d)$. On suppose que*

$$\mathbb{P}[\boldsymbol{X} \in \mathcal{L}_D(\alpha)] > 0.$$

*Sous l'hypothèse **(H1)**, et en supposant qu'il existe un $r \geq 2$ tel que Y est r-intégrable, on a :*

$$n^{\frac{r-1}{2r}} \wedge v_n^{\frac{(p-1)(r-1)}{pr}} \left| \widehat{\mathrm{CCTE}}_{D,\alpha}^{n}(Y,\boldsymbol{X}) - \mathrm{CCTE}_{D,\alpha}(Y,\boldsymbol{X}) \right| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

En contrôlant la distance de Haussdorf entre $\mathcal{L}_n$ et $\mathcal{L}_D$ on est également capable de lier la vitesse de convergence de la $\widehat{\text{CCTE}}$ en fonction de celle d'un estimateur D_n de D (théorème ci-dessous).

Théorème 3. *Sous ceratines hypothèses techniques sur la profondeur D (plus précisemment, sur son comportement au voisinage du niveau α ainsi que sa consistance) on a,*

$$\lambda_d(\mathcal{L}(\alpha)\Delta\mathcal{L}_n(\alpha)) = \underset{n\to\infty}{O}(\|D_n - D\|_{\infty,\mathbb{R}^d}),\ \mathbb{P}\text{-}p.s.$$

Par conséquent, il s'agit d'étudier la vitesse de convergence de $\|D_n - D\|_\infty$ vers zéro en probabilité. Ceci nous permettra de conclure sur la vitesse de convergence de la CCTE. Notre objectif final sera donc de proposer des profondeurs et leurs estimateurs permettant d'estimer efficacement notre CCTE. Enfin, nous proposerons une étude sur simulation pour illustrer nos résultats.

Références

F. Belzunce, A. Castaño, A. Olvera-Cervantes, and A. Suárez-Llorens. Quantile curves and dependence structure for bivariate distributions. *Computational Statistics & Data Analysis*, 51(10) :5112–5129, 2007.

F. Chebana and T. BMJ Ouarda. Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22(1) :63–78, 2011.

A. Cousin and E. Di Bernardino. On multivariate extensions of value-at-risk. *Journal of multivariate analysis*, 119 :32–46, 2013.

L. Dehaan and X. Huang. Large quantile estimation in a multivariate setting. *Journal of Multivariate Analysis*, 53(2) :247–263, 1995.

T. Laloë, R. Servien, and E. Di Bernardino. Estimating covariate functions associated to multivariate risks : a level set approach. *Metrika, Springer Verlag*, pages 497–526, 2015.

R. Torres, E. Di Bernardino, H. Laniado, and R. Lillo. On the estimation of extreme directional multivariate quantiles. *Communications in Statistics-Theory and Methods*, 49(22) :5504–5534, 2020.

Y. Zuo and R. Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

Analyse de la pertinence des couches dans les architectures de forêts profondes

Ludovic Arnould [1], Claire Boyer [1] & Erwan Scornet [2]

[1] *LPSM, Sorbonne Université, ludovic.arnould@sorbonne-universite.fr*
claire.boyer@sorbonne-universite.fr
[2] *CMAP, Ecole Polytechnique, erwan.scornet@polytechnique.edu*

Résumé. Les forêts aléatoires, d'une part, et les réseaux de neurones, d'autre part, ont rencontré un grand succès dans la communauté de l'apprentissage machine pour leurs performances prédictives. Des combinaisons des deux approches ont été proposées dans la littérature, conduisant notamment aux forêts dites profondes (DF) (Zhou & Feng,2019). Dans ce travail, notre objectif n'est pas de comparer les performances des forêts profondes avec celles d'autres méthodes, mais d'étudier les mécanismes sous-jacents de ce modèle. En outre, nous montrons que l'architecture des forêts profondes peut généralement être simplifiée en réseaux de forêts peu profonds, plus simples et plus efficaces sur le plan algorithmique. Malgré une certaine instabilité, ces derniers peuvent surpasser les méthodes prédictives standard reposant sur des arbres de décision. Nous étudions théoriquement une architecture simplifiée : nous présentons un cadre théorique dans lequel un réseau d'arbres peu profond améliore les performances des arbres de décision classiques. Pour ce faire, nous établissons des bornes théoriques sur le risque pour des arbres classiques et des réseaux d'arbres. Ces bornes révèlent l'intérêt des architectures de réseaux d'arbres lorsque les données sont très structurées, et à condition que la première couche, servant d'encodeur, soit suffisamment riche.

Mots-clés : Forêts profondes, apprentissage profond, forêts aleatoires, réseaux de neurones, modèles hybrides, apprentissage statistique.

Abstract. Random forests on the one hand, and neural networks on the other hand, have met great success in the machine learning community for their predictive performance. Combinations of both have been proposed in the literature, notably leading to the so-called deep forests (DF) (Zhou & Feng,2019). In this paper, our aim is not to benchmark DF performances but to investigate instead their underlying mechanisms. Additionally, DF architecture can be generally simplified into more simple and computationally efficient shallow forest networks. Despite some instability, the latter may outperform standard predictive tree-based methods. We theoretically study a simplified architecture: we exhibit a theoretical framework in which a shallow tree network is shown to enhance the performance of classical decision trees. In such a setting, we provide tight theoretical lower and upper bounds on its excess risk. These theoretical results show the interest of tree-network architectures for well-structured data provided that the first layer, acting as a data encoder, is rich enough.

Keywords: Deep Forest, deep learning, random forests, neural network, hybrid models, statistical learning.

1 Introduction

Récemment, plusieurs approches ont été proposées pour envisager des réseaux de neurones (DNNs) avec des modules non différentiables. Parmi elles, l'algorithme Deep Forest **?**, qui utilise des forêts aléatoires (RF pour "Random Forest") en guise de neurones, a reçu beaucoup d'attention ces dernières années dans diverses applications telles que le traitement d'images hyperspectrales **?**, l'imagerie médicale **?**, les interactions médicamenteuses **??** ou encore la détection de fraudes **?**.

Les forêts profondes (DF pour "Deep Forest") **?** consistent en une procédure hybride dans laquelle des forêts aléatoires sont utilisées comme composants élémentaires (neurones) d'un réseau de neurones. Chaque couche de DF est composée d'un assortiment de forêts de Breiman et de forêts complètement aléatoires (CRF) **?**, entraînées les unes à la suite des autres, voir Figure 1.

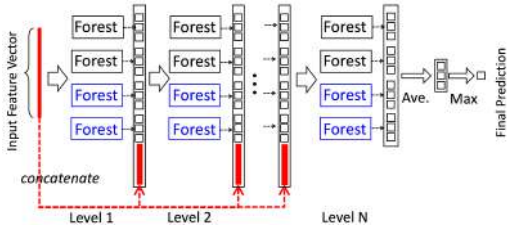


Figure 1: Architecture des forêts profondes (le schéma provient de **?**). Le vecteur de sortie de chaque couche est concaténé avec les entrées avant d'être passés à la couche suivante.

Les méthodes DF présentent de bonnes performances en pratique, ce qui suggère que l'empilement de RF et l'extraction des caractéristiques de ces estimateurs à chaque couche du réseau est un moyen prometteur de tirer parti de leurs performances dans le cadre des réseaux neuronaux. Les travaux afférents aux DF représentent cependant exclusivement des contributions algorithmiques sans compréhension formelle des mécanismes moteurs à l'œuvre dans les cascades de forêts. La procédure reste en effet complexe : plusieurs couches sont empilées, chacune étant composée d'estimateurs RF non paramétriques, eux-mêmes complexes. Les raisons du gain de performance accordé par de telles architectures restent donc opaques.

L'objectif de ce travail n'est pas de faire une étude exhaustive des performances de prédiction des DFs **?** mais plutôt de comprendre comment l'empilement d'arbres en réseau peut aboutir à une procédure d'apprentissage compétitive.

Contributions. Dans cette étude, nous analysons l'avantage de combiner les arbres dans une architecture réseau à la fois théoriquement et numériquement. Nos principaux objectifs sont (i) de quantifier les avantages potentiels de la DF par rapport à la RF, et (ii) de comprendre les mécanismes à l'œuvre dans des architectures aussi complexes. Nous montrons en particulier qu'une configuration peu profonde peut aboutir aux mêmes performances que la configuration par défaut des DF. Nous montrons aussi que la performance de la méthode globale dépend essentiellement de la structure des premières couches. Pour une architecture simplifiée, et dans un cadre théorique supposant une certaine structure des données, nous établissons des bornes inférieures et supérieures pour les risques d'un arbre et d'un réseau d'arbres. Nous prouvons ainsi qu'un réseau d'arbres peu profond peut être plus performant qu'un arbre individuel dans le cas spécifique de données structurées si le premier arbre du réseau joue le rôle d'un encodeur efficace.

2 Analyse numérique des architectures DF

Une grande profondeur est-elle nécessaire ? Nous éprouvons numériquement les DF avec différents nombres de couches sur des jeux de données réels : une version simplifiée des DF à peu de couches donne des résultats équivalents à ceux d'une DF profonde, tout en permettant une réduction drastique du nombre de paramètres des méthodes. Pour la plupart des jeux de données, envisager les DF à deux couches constitue déjà une amélioration par rapport à l'algorithme RF de base.

A la recherche du meilleur sous-modèle. Nous conduisons ensuite une analyse approfondie du rôle de chaque couche dans la performance globale des DF : dans une architecture à L couches, nous pointons, pour chaque expérience, la couche offrant les meilleures prédictions. Globalement, sur la plupart des jeux de données, nous observons que les sous-modèles à une ou deux couches donnent les meilleures performances.

La meilleure performance des premiers sous-modèles peut s'expliquer par la pauvreté de la nouvelle représentation créée par chaque couche : dans un cadre de classification multi-label, la dimension du nouveau vecteur de variables créées correspond au nombre de classes multiplié par le nombre de forêts qui peut être faible, suivant l'architecture, devant le nombre de caractéristiques d'entrée ; d'autre part, les différentes forêts au sein d'une couche sont susceptibles de produire des sorties similaires.

Compréhension précise de l'influence de l'encodage dans un réseau d'arbres à deux couches. Afin d'appréhender finement l'influence de la profondeur des arbres en DF, nous étudions une architecture simplifiée : un réseau d'arbres CART, composé de deux couches comptant un seul arbre CART par couche.

Pour chaque échantillon, l'arbre de première couche produit un vecteur de probabilités (ou un scalaire dans un cadre de régression), que l'on peut assimiler à un "encodage" des

données d'entrée et qui est passé à l'arbre de la seconde couche avec les données brutes.

On observe que lorsque l'encodage est de mauvaise qualité (i.e. l'arbre de la première couche est de profondeur sous-optimale), le second arbre ne peut pas améliorer les résultats du premier arbre tant qu'il n'est pas assez développé pour contrer l'effet de l'encodage.

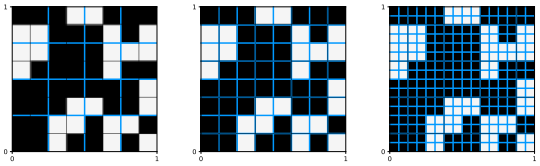
3 Etude théorique d'un réseau d'arbres peu profond

Dans cette section, nous nous concentrons sur l'analyse théorique d'une architecture simplifiée : notre objectif est de mettre en exergue des situations dans lesquelles un réseau d'arbres à deux couches est plus performant qu'un arbre individuel. Parce que la deuxième couche d'un réseau d'arbres permet de rassembler les feuilles d'arbres de la première couche avec des distributions similaires, utiliser un réseau d'arbres semble judicieux lorsque l'ensemble de données a une structure très spécifique, dans laquelle le même lien entre l'entrée et la sortie peut être observé dans différentes sous-parties de l'espace d'entrée.

Architecture étudiée. Nous nous concentrons sur **deux arbres centrés en cascade** (voir des partitions correspondantes sur Figure 2) et nous tentons de déterminer l'influence du premier arbre (d'encodage) sur les performances de l'ensemble du réseau d'arbres. Le prédicteur résultant composé des deux arbres en réseau, de profondeurs respectives k et k', formés sur les données $(X_1, Y_1), \ldots, (X_n, Y_n)$ est désigné par $\hat{r}_{k,k',n}$. Soit $r(x) = \mathbb{E}[Y|X = x]$ la fonction de régression à estimer et pour toute fonction f, le risque quadratique associé est noté $R(f)$.

Distribution des données. On considère un échantillon $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ de copies i.i.d. d'un couple générique (X, Y) avec $X \sim \mathcal{U}([0,1]^2)$ et $Y \in \{0, 1\}$. L'espace d'entrée $[0,1]^2$ est arbitrairement découpé en cellules blanches et noires, cf Figure 2. Si X tombe dans une case noire (resp. blanche) Y suit alors une loi de Bernoulli de paramètre p (resp. $1-p$). Ainsi, la distribution des données est parametrée par $k^\star$ (avec $2^{k^\star}$ le nombre total de cellules), p et $N_{\mathcal{B}}$ le nombre de cellules noires (cf. Figure 2). La fonction de régression correspondante est donc constante par morceaux sur $[0,1]^2$ prenant deux valeurs possibles p ou $1-p$.

Principaux résultats. Nous étudions la performance d'un arbre individuel vs. celle d'un réseau d'arbres pour l'estimation de la fonction de régression précédente. Nous distinguons essentiellement deux cas : lorsque la profondeur du premier arbre est sous-optimale ($k < k^\star$) ou bien lorsque le premier arbre est assez développé ($k \geq k^\star$). Pour le premier cas, le biais du premier arbre sera maximal lorsque la distribution des données comptera autant de cases noires que de cases blanches réparties de manière régulière, i.e. dans le cas d'une distribution structurée suivant un damier régulier.



(a) Depth 4 (b) Depth 6 (c) Depth 8

Figure 2: Distribution des données selon un damier "généralisé" avec $k^\star = 6$ et $N_{\mathcal{B}} = 40$ cellules noires . La partition du (premier) arbre d'encodage de profondeur 4, 6 et 8 (de gauche à droite) est affichée en bleu. La profondeur optimale d'un seul arbre centré pour cette structure de données est de 6 (milieu).

Proposition 1 (Bornes sur le risque quand $k < k^\star$)**.** *Ici nous supposons que les données sont tirées selon un damier régulier contenant autant de cases noires et blanches, de paramètres $k^\star$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ et $p > 1/2$. Dans ce cas, chaque feuille du premier arbre contient autant de cellules blanches que noires et la prédiction dans chaque feuille est proche de $1/2$ en moyenne. Le second arbre est alors aussi biaisé (à moins de le développer jusqu'à la profondeur optimale $k^\star$).*

Proposition 2 (Bornes sur le risque quand $k \geq k^\star$)**.** *On considère un damier généralisé (avec un nombre et un positionnement arbitraires des cellules noires) de paramètres $k^\star$, $N_{\mathcal{B}}$ et $p > 1/2$.*

1. *Soit un arbre simple $\hat{r}_{k,0,n}$ de profondeur $k \in \mathbb{N}^\star$. On a*

$$R(\hat{r}_{k,0,n}) \leq \frac{2^k p(1-p)}{n+1} + \left(p^2 + (1-p)^2\right) \frac{(1-2^{-k})^n}{2},$$

et

$$R(\hat{r}_{k,0,n}) \geq \frac{2^{k-1} p(1-p)}{n+1} + \left(p^2 + (1-p)^2 - \frac{2^k p(1-p)}{n+1}\right) \frac{(1-2^{-k})^n}{2}.$$

2. *On considère un réseau d'arbre $\hat{r}_{k,1,n}$. On a*

$$R(\hat{r}_{k,1,n}) \leq 2 \cdot \frac{p(1-p)}{n+1} + \frac{2^{k+1} \varepsilon_{n,k,p}}{n} + \left(\frac{N_{\mathcal{B}}}{2^{k^\star}} p^2 + \frac{2^{k^\star} - N_{\mathcal{B}}}{2^{k^\star}} (1-p)^2\right) (1-2^{-k})^n$$

où $\varepsilon_{n,k,p} = n(1 - \frac{1 - e^{-2(p-\frac{1}{2})^2}}{2^k})^n$.

De plus la borne inférieure est du même ordre de grandeur que la borne supérieure.

D'après cette proposition, lorsque le premier arbre n'est pas biaisé, le second arbre à l'aide d'une seule coupe sert de réducteur de variance, améliorant le risque global d'un facteur 2^k comparé à celui d'un arbre individuel. Notons que cette réduction de la variance ne peut pas être obtenue en faisant la moyenne de nombreux arbres, comme dans une structure standard de forêts aléatoires centrées. Cela montre l'avantage des architectures à couches d'arbres par rapport aux méthodes d'ensemble classiques.

Bibliographie

[1] Wei Fan, Haixun Wang, Philip S Yu, and Sheng Ma. Is random model better? onits accuracy and efficiency. InThird IEEE International Conference on Data Mining,pages 51–58. IEEE, 2003.

[2] B. Liu, W. Guo, X. Chen, K. Gao, X. Zuo, R. Wang, and A. Yu. Morphologicalattribute profile cube and deep random forest for small sample classification of hyperspectral image.IEEE Access, 8:117096–117108, 2020.

[3] R. Su, X. Liu, L. Wei, and Q. Zou. Deep-resp-forest: A deep forest model to predictanti-cancer drug response.Methods, 166:91–102, 2019.

[4] L. Sun, Z. Mo, F. Yan, L. Xia, F. Shan, Z. Ding, B. Song, W. Gao, W. Shao, F. Shi,H. Yuan, H. Jiang, D. Wu, Y. Wei, Y. Gao, H. Sui, D. Zhang, and D. Shen. Adaptivefeature selection guided deep forest for covid-19 classification with chest ct.IEEEJournal of Biomedical and Health Informatics, 24(10):2798–2805, 2020.

[5] X. Zeng, S. Zhu, Y. Hou, P. Zhang, L. Li, J. Li, L F. Huang, S. J Lewis, R. Nussi-nov, and F. Cheng. Network-based prediction of drug–target interactions using anarbitrary-order proximity embedded deep forest.Bioinformatics, 36(9):2805–2812,2020.

[6] Y. Zhang, J. Zhou, W. Zheng, J. Feng, L. Li, Z. Liu, M. Li, Z. Zhang, C. Chen, X. Li,et al. Distributed deep forest and its application to automatic detection of cash-outfraud.ACM Transactions on Intelligent Systems and Technology (TIST), 10(5):1–19,2019.

[7] Z. Zhou and J. Feng. Deep forest.National Science Review, 6(1):74–86, 2019.

Omic Fold changes clustering to study the radiation response of endothelial cells

Polina Arsenteva [1] & Vincent Paget [2] & Olivier Guipaud [3] & Fabien Milliat [4] & Hervé Cardot [5] & Mohamed Amine Benadjaoud [6]

[1,5] *IMB, Université de Bourgogne. 9 avenue Alain Savary, 21078 Dijon Cedex*
[2,3,4] *IRSN, PSE-SANTE/SERAMED/LRMed.*
[6] *IRSN, PSE-SANTE/SERAMED.*
31 avenue de la Division Leclerc, 92262 Fontenay-aux-Roses Cedex
[1] *polina.arsenteva@u-bourgogne.fr* [2] *vincent.paget@irsn.fr*
[3] *olivier.guipaud@irsn.fr* [4] *fabien.milliat@irsn.fr*
[5] *herve.cardot@u-bourgogne.fr* [6] *mohamedamine.benadjaoud@irsn.fr*

Résumé. En radiobiologie, un "fold change" est défini comme une mesure de différence entre l'état irradié et non-irradié d'une certaine entité biologique au cours du temps. L'objectif de ce travail est de proposer une méthode de clustering des fold changes afin d'identifier les principales modalités de réponse de cellules endothéliales suite à une irradiation in vitro à travers différents omiques: gènes, protéines et métabolites. Notre approche est basée sur le clustering des lois de probabilité des fold changes qui sont supposées gaussiennes. Nous proposons une modification de la méthode de clustering des lois basée sur la distance de Wasserstein tout en tenant compte des corrélations entre les variables estimées à partir des données omiques. La performance de la nouvelle méthode est évaluée et comparée à celle de la méthode de clustering k-means basée sur la distance de Wasserstein.

Mots-clés. Irradiation in vitro, données omiques, analyse des données fonctionnelles, clustering, distance de Wasserstein.

Abstract. In radiobiology, a fold change is defined as a measure of difference between irradiated and non-irradiated condition for a certain biological entity over time. The goal of this work is to propose a fold change clustering method in order to identify major response modalities of endothelial cells after in vitro radiation exposure through transcriptomic, proteomic and metabolomic data. Our approach is based on clustering fold changes' probability distributions that are assumed to be Gaussian. We propose a modification of distribution clustering method based on Wasserstein distance while taking into account correlations between variables estimated from an omics data set. Performance of the new method is evaluated and compared to the baseline Wasserstein distance-based k-means clustering.

Keywords. In vitro irradiation, omics data, functional data analysis, clustering, Wasserstein distance.

1 Introduction

More than 200 000 patients undergo radiotherapy in France every year. This treatment, however, is associated with undesirable effects for the healthy tissues situated in close proximity to the irradiated tumors. It is thus of interest to compare different radiation treatment configurations characterized by differences in dose, volume, energy, etc. in order to ultimately suggest a treatment associated with minimal risks for patients.

The study of radiation response can be split into different scales, and this work focuses on the radiation response of endothelial cells, previously identified as a key actor in the appearance of radiation adverse effects. To investigate this cellular response, multiple omics data sets (transcriptomic measuring gene expression, proteomic for protein expression, etc.) were collected for several time points. The common feature of all data sets is the presence of two experimental conditions: irradiated and non-irradiated (control). The quantity of interest that will be distinguished is radio-induced fold change: a measure of irradiation effect represented by the difference between the two experimental conditions over time.

In the course of this project, functional ANOVA model is used to estimate the fold changes as temporal curves. These estimations are thus subject to uncertainties that should be considered in subsequent investigations, such as clustering. The goal is then to deduce specific temporal characteristics of the fold changes for different biological metrics, in particular to determine the most characteristic time points that will be used for further experiments. Since the behavior shown by the curves appears to be extremely variable, it is reasonable to perform data clustering as means of distinguishing multiple characteristic types of omics radiation responses. In this preliminary work, we introduce a distribution-based clustering problem starting from the classical Wasserstein distance, and propose an alternative algorithm performing temporal curve clustering in the presence of uncertainties and joint correlations.

2 Functional ANOVA model

Consider an observation from one of the studied data sets characterized by the following quantities: $i \in \{1, 2, ..., N\}$ where N is the number of considered biological entities, replicate $j \in \{1, 2, ..., n_r\}$, and experimental condition $k = 0$ if control and $k = 1$ if irradiated. Here the example of transcriptomic data set will be given, in this case i refers to a specific gene, N is the number of genes present in the study, and $n_r = 3$. Let $y_{ikj}(t)$ be a realization of the expression of a gene i under experimental condition k for replicate j at time point $t \in \mathbb{R}^+$. We consider the following model :

$$y_{ikj}(t) = \mu(t) + \mathbb{1}_{k=1}\alpha(t) + \beta_{ik}(t) + \epsilon_{ikj}(t). \tag{1}$$

The variables appearing in the model are as follows : $\mu(t)$ is the grand mean function,

$\alpha(t)$ is the global fold change (that is, present in all genes), $\beta_{ik}(t)$ is the individual gene effect. We can thus define local fold change for a gene i: $\Delta_i(t) = \beta_{i1}(t) - \beta_{i0}(t)$.

This model corresponds to a two-way functional ANOVA (f-ANOVA) model [1], with the general form presented below:

$$y_{ikj}(t) = \mu(t) + \alpha_k(t) + \beta_i(t) + \theta_{ik}(t) + \epsilon_{ikj}(t). \tag{2}$$

In this model we can distinguish the main-effect of irradiation $\alpha_k(t) = \mathbb{1}_{k=1}\alpha(t)$, the main-effect of a gene $\beta_i(t) = \beta_{i0}(t)$, and the interaction-effect between irradiation and a gene: $\theta_{ik}(t) = 0$ if $k = 0$ and $\theta_{ik}(t) = \Delta_i(t)$ if $k = 1$.

The model can be rewritten in the following form allowing for direct estimation of regression parameters as presented in [2]:

$$Y(t) = Z\gamma(t) + \epsilon(t), \tag{3}$$

with $Y(t) = \begin{pmatrix} y_{101}(t) \\ y_{102}(t) \\ y_{103}(t) \\ y_{111}(t) \\ y_{112}(t) \\ y_{113}(t) \\ y_{201}(t) \\ \vdots \\ y_{N13}(t) \end{pmatrix}$, $\gamma(t) = \begin{pmatrix} \mu(t) \\ \alpha(t) \\ \beta_{10}(t) \\ \beta_{11}(t) \\ \vdots \\ \beta_{N0}(t) \\ \beta_{N1}(t) \end{pmatrix}$ and $Z = \begin{pmatrix} 1 & \cdots & 1 \\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 1 & \cdots & 0\ 1\ 1\ 1 \\ 1\ 1\ 1\ 0 & \cdots & 0 \\ 0\ 0\ 0\ 1\ 1\ 1\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0\ 1\ 1\ 1\ 0\ 0\ 0 \\ 0 & \cdots & 0\ 1\ 1\ 1 \end{pmatrix}^t$.

3 Fold Change Clustering

The task of identification of different groups of genes with respect to their response to irradiation can be translated into that of clustering fold changes of genes. A fold change of a gene i will be denoted as $\Gamma_i = (\Gamma_i(t_1), ..., \Gamma_i(t_p)), p \in \mathbb{N}^*$. With respect to the f-ANOVA model described in the previous section, a fold change of a gene is defined as a sum of a global and a local fold change: $\Gamma_i(t) = \alpha(t) + \Delta_i(t)$. In this section time is discrete and $t_1, ..., t_p$ are the observed time points. Fold changes are estimated from the data for p time points with associated uncertainties. In order to take into account the uncertainty of this estimation, the fold changes are considered to be normally distributed: $\Gamma_i \sim \mathcal{N}(\mu_{\Gamma_i}, \Sigma_{\Gamma_i})$. The task is therefore to perform distribution clustering which requires choosing an appropriate distance.

Associated with a number of appealing properties, the Wasserstein distance allows to compare probability distributions and is defined as follows:

$$W_2^2(P_1, P_2) = \inf_J \int \|x - y\|^2 dJ(x, y), \tag{4}$$

where J is a joint distribution of random variables with marginals P_1 and P_2. In other words, the Wasserstein distance between probability distributions P_1 and P_1 is the quantity $\mathbb{E}\|X - Y\|^2$ (where X and Y are random variables such that $X \sim P_1$ and $Y \sim P_2$)

minimized over the set of all possible joint distributions of X and Y (for details see [3]). The benchmark goal is to perform a Wasserstein distance-based k-means clustering of distributions of fold changes $\Gamma = (\Gamma_1, ..., \Gamma_N)$.

On the one hand, from the biological perspective it is justified to assume that some groups of genes are mutually dependent, which is why it is of interest to take the correlations between genes into account when performing clustering. On the other hand, the presence of multiple replicates in the data set allows to estimate the joint distributions of all variable pairs representing a fold change while taking into account the correlations between genes, i.e. $\mathrm{Cov}(y_{ikj}(t), y_{i'kj}(t)) \neq 0$ for $i \neq i'$. For these reasons, the following distance will be considered:

$$d_2^2(P_1, P_2) = d_2^2(X, Y) = \mathbb{E}\|X - Y\|^2. \tag{5}$$

Instead of minimizing the expected value of the norm over the set of all possible joint distributions, the latter is calculated for the actual joint distribution, which is considered to be known. In particular, for Gaussian variables $\Gamma_i \sim \mathcal{N}(\mu_{\Gamma_i}, \Sigma_{\Gamma_i})$ and $\Gamma_j \sim \mathcal{N}(\mu_{\Gamma_j}, \Sigma_{\Gamma_j})$, the distance can be developed in the following way:

$$d_2^2(\Gamma_i, \Gamma_j) = \mathbb{E}\|\Gamma_i - \Gamma_j\|^2 = \|\mu_{\Gamma_i} - \mu_{\Gamma_j}\|^2 + \mathrm{Tr}(\Sigma_{\Gamma_i}) + \mathrm{Tr}(\Sigma_{\Gamma_j}) - 2\mathrm{Tr}(K), \tag{6}$$

where K is such that appears in the covariance matrix of $\Gamma_{ij} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{\Gamma_i}\\ \mu_{\Gamma_j}\end{bmatrix}, \begin{bmatrix}\Sigma_{\Gamma_i} & K\\ K^T & \Sigma_{\Gamma_j}\end{bmatrix}\right)$, the joint distribution of Γ_i and Γ_j. Covariance matrices $\Sigma_{\Gamma_i} = (\sigma^2_{\Gamma_i}(t_l, t_m)) \in \mathbb{R}^{p\times p}$, $\Sigma_{\Gamma_j} = (\sigma^2_{\Gamma_j}(t_l, t_m)) \in \mathbb{R}^{p\times p}$ and $K = (\rho_{\Gamma_i\Gamma_j}(t_l, t_m)) \in \mathbb{R}^{p\times p}$ are assumed to be symmetric positive definite and are initially considered as diagonal due to the following characteristic of the experimental setting and hence the data set: the in vitro extraction of omic expressions is a destructive process, which implies that a measure at each time point is performed on a different pool of endothelial cells. Thus, the measures at two distinct time points could be considered as independent. This assumption will be relaxed in future investigations in order to include more biologically realistic temporal correlations. The distance estimate that will be used in clustering can then be written explicitly:

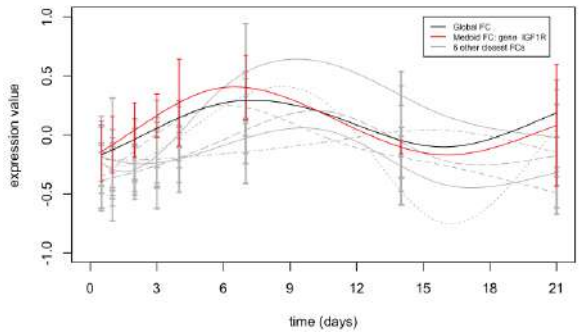
$$\widehat{d_2^2(\Gamma_i, \Gamma_j)} = \sum_{l=1}^{p} (\widehat{\mu_{\Gamma_i(t_l)}} - \widehat{\mu_{\Gamma_j(t_l)}})^2 + \sum_{l=1}^{p} \widehat{\sigma^2_{\Gamma_i(t_l)}} + \sum_{l=1}^{p} \widehat{\sigma^2_{\Gamma_j(t_l)}} - 2\sum_{l=1}^{p} \widehat{\rho_{\Gamma_i\Gamma_j(t_l)}}. \tag{7}$$

Unfortunately, it is impossible to combine the k-means approach with the distance estimate presented above. As soon as the first k barycenters are calculated, the distances that have to be considered are no longer those between the pairs of observed fold changes, but those between every fold change and the barycenters associated with every one of the k clusters. These barycenters are not observed, their distributions are Gaussian and represented only by a mean vector and a covariance matrix each. The joint distributions are thus inaccessible, making it impossible to calculate the necessary distances. We will thereupon concentrate on implementing the following algorithms:

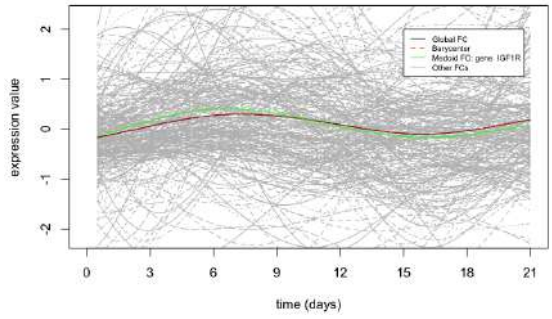
- **Wasserstein distance-based k-means:** this option can be considered as a benchmark, while keeping in mind that in this case the correlations between genes are not taken into account.
- $\widehat{\mathbf{d_2^2}}$**-based k-medoids:** the previously mentioned problem can be solved by replacing barycenters with medoids, that is the observed fold changes that are the closest to every fold change in the given cluster. In this case the only needed distances are those between the pairs of observed fold changes, providing access to the joint distributions' estimates. Hence, we can use the distance based on the distribution estimate for clustering while taking into account the correlations between genes.

The first option was implemented according to the algorithm described in [3]. In the implementation of the k-medoids version, the initialization step was performed in the same way, the rest of the algorithm was implemented according to Partitioning Around Medoids (PAM) algorithm (e.g. see [4]). A slight modification was introduced with the goal of improving performance.

4 Preliminary results & Further work



(a) Functional representations of 7 fold changes closest to the global fold change (their medoid in red).



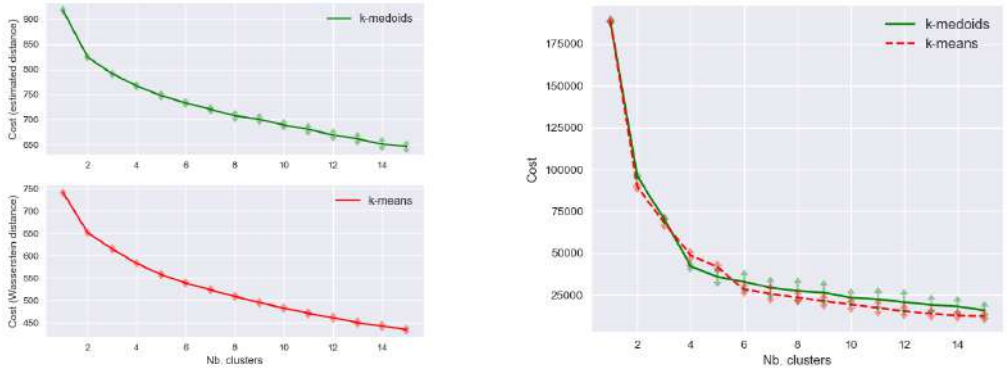
(b) Comparison of functional representations of the medoid (k-medoids) and the barycenter (k-means) with the global fold change in the case of single cluster.

Figure (1)

We have tested the algorithms on the transcriptomic data set and compared results. The centroids in the case of a single cluster have been studied for the purpose of testing the performance of the medoid search part of k-medoids algorithm as well the barycenter search part of k-means algorithm. Both the medoid and the barycenter are expected to be close to the global fold change α. Functional versions of fold changes including the medoid curve and the barycenter were compared. The medoid detected by the k-medoids version of the algorithm is very close to the global fold change, it appears to be the closest one out of all fold changes (see Figure 1a). The barycenter detected by the k-means version of the algorithm is exactly the same as the global fold change (see Figure 1b). It can be concluded that the barycenter is a better approximation of the global fold change, which is also expected since the medoid is chosen among the existing gene fold changes whereas the

barycenter is estimated without this constraint. All of the previous arguments considered, it can be concluded that both algorithms successfully perform the part of the clustering procedure that corresponds to updating centroids for given clusters.

Some preliminary comparative results in case of multiple clusters are available. Figure 2a presents costs functions that are being optimized by each algorithm demonstrating steady decline with the growing number of clusters. A cost function summing pairwise distances $\widehat{\mathbf{d_2^2}}$ within every cluster was used to compare clustering success for different numbers of clusters. Figure 2b shows that an elbow for this cost function in case of k-medoids algorithm appears at 4 clusters whereas in case of k-means at 6 clusters. Obtaining fewer clusters in the first case is consistent with the idea of taking correlations into account and as a result identifying redundant information.



(a) Native cost functions for both methods. (b) Costs comparison.

Figure (2)

Currently various simulation studies are conducted in order to test and compare the robustness and the sensitivity of the two clustering techniques by introducing several levels of correlations in the joint distributions at a given time point. The next step of this case study is to estimate covariances with respect to time based on a markovian model, incorporate them into the covariance matrices and perform fold change clustering while accounting for all possible variable correlations.

References

[1] Jin-Ting Zhang. (2013). *Analysis of Variance for Functional data.* Chapman & Hall Book; Press, C.R.C., Ed.; Taylor & Francis Group: Abingdon, UK.

[2] James O. Ramsay, Bernard Silverman. (2005). *Functional Data Analysis.* Springer Science & Business Media.

[3] Isabella Verdinelli, Larry Wasserman. (2019). *Hybrid Wasserstein distance and fast distribution clustering.* Electron. J. Statist. 13, no. 2, 5088–5119.

[4] Erich Schubert, Peter J. Rousseeuw. (2019). *Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms.* Similarity Search and Applications, Springer International Publishing, 11807.

Estimation alternative des paramètres d'un mélange de régressions binaires

Benjamin Auder [1] & Élisabeth Gassiat [2] Mor-Absa Loum [3]

[1] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, benjamin.auder@universite-paris-saclay.fr* [2] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, elisabeth.gassiat@universite-paris-saclay.fr* [3] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, morabsa.loum5@gmail.com*

Résumé. Considérons un mélange de modèles de régression linéaire généralisée à sorties binaires : $\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle x, \beta_k \rangle + b_k)$, dont les paramètres sont traditionnellement estimés en maximisant la vraisemblance. Nous proposons une estimation alternative basée sur l'adéquation des moments croisés empiriques avec leurs analogues théoriques. L'identitifiabilité, la consistance ainsi que la normalité asymptotique sont démontrées. Les simulations effectuées sont très encourageantes quant à l'intérêt pratique de la méthode, implémentée dans un package R disponible sur le CRAN.

Mots-clés. Mélange, régression binaire, moments

Abstract. We consider finite mixtures of generalized linear models with binary output: $\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle x, \beta_k \rangle + b_k)$, which parameters are usually estimated by maximizing the likelihood. We propose an alternative estimation based on the adequation of the empirical cross-moments with their theoretical counterparts. Identifiability, consistency as well as asymptotic normality are proved. Simulations run are very promising concerning the practical interest of the method, implemented in a R package available on CRAN.

Keywords. Mixture, binary regression, moments

1 Introduction

Le modèle de régression logistique (ou de régression linéaire généralisée) est très populaire dans divers domaines. Quand les données étudiées proviennent de plusieurs groupes latents, utiliser des modèles de mélanges est une manière courante de gérer l'hétérogénéité. Beaucoup d'algorithmes ont ainsi été développés pour estimer les paramètres de tels modèles, la plupart basés sur la maximisation de la (log-)vraisemblance via un algorithme E-M. L'objectif de ce travail est d'explorer une approche alternative d'estimation des paramètres basée sur les moments croisés jusqu'à l'ordre 3, dans le cadre des méthodes basées sur les tenseurs − voir par exemple Anandkumar et al (2014).

Soit $X \in \mathbb{R}^d$ le vecteur des covariables et $Y \in \{0,1\}$ la sortie binaire. Selon un modèle de régression binaire, la probabilité conditionnelle $\mathbb{P}(Y = 1|X = x)$ est donnée par $g(\langle \beta, x\rangle + b)$, où $\beta \in \mathbb{R}^d$ est le vecteur des coefficients de régression, $b \in \mathbb{R}$ désignant l'ordonnée en zéro. Le package que nous avons développé permet d'utiliser les liens logit et probit, où g est donnée respectivement par $g(z) = e^z/(1+e^z)$ et $g(z) = \Phi(z)$ avec Φ la fonction de répartition de la loi gaussienne standard $\mathcal{N}(0,1)$.

Si maintenant l'on souhaite modéliser des populations hétérogènes, fixons K le nombre de populations et $\omega = (\omega_1, \cdots, \omega_K)$ leurs poids tels que $\omega_j \geq 0$, $j = 1 \ldots, K$ et $\sum_{j=1}^K \omega_j = 1$. L'expression précédente se généralise naturellement en

$$\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle \beta_k, x\rangle + b)\,.$$

Notons $\theta = (\omega, \beta, b)$ l'ensemble des paramètres.

Trois résultats théoriques sont démontrés dans Auder et al (2020) : l'identifiabilité (à partir des moments croisés), la convergence et la normalité asymptotique. Cependant, nous préférons ici nous concentrer sur les aspects pratiques de la méthode d'estimation, décrite ci-après. Afin de simplifier les choses et sans perdre de généralité, la covariable X sera supposée suivre une loi gaussienne standard : $X \sim \mathcal{N}(0,1)$. Les résultats peuvent s'étendre à dautre distributions, mais ce n'est pas l'objet de cette communication.

2 Méthode d'estimation

2.1 Idée générale

Commençons par définir les moments croisés : en notant e_j le j^{eme} vecteur de la base canonique de $\mathbb{R}^d$, ceux-ci s'écrivent ainsi :

- $M_1(\theta) := E_\theta[YX]$,
- $M_2(\theta) := E_\theta\left[Y\left(X \otimes X - \sum_{j\in[d]} e_j \otimes e_j\right)\right]$, and
- $M_3(\theta) := E_\theta\left[Y\left(X \otimes X \otimes X - \sum_{j\in[d]} [X \otimes e_j \otimes e_j + e_j \otimes X \otimes e_j + e_j \otimes e_j \otimes X]\right)\right]$.

Le produit tensoriel s'effectue composante par composante, généralisant le produit matriciel. Par exemple $e_j \otimes e_j$ est une matrice et $X \otimes e_j \otimes e_j$ un tenseur d'ordre 3.

Seuls les moments basés sur la vraie valeur du paramètre ($\theta^\star$) sont utiles. Mais comme ce paramètre est inconnu, on approche empiriquement les vrais moments croisés :

$$\begin{aligned}
\widehat{M_1} &= \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \\
\widehat{M_2} &= \frac{1}{n}\sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i - \sum_{j\in[d]} e_j \otimes e_j) \Big] \\
\widehat{M_3} &= \frac{1}{n}\sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i \otimes X_i - \sum_{j\in[d]} \big[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i \big]) \Big].
\end{aligned}$$

D'autres part, les moments théoriques peuvent s'exprimer simplement à partir de θ :

$$\begin{aligned}
M_1(\theta) &= \sum_{k=1}^{K} \omega_k E[g'\big(\langle X, \beta_k\rangle + b_k\big)]\ \beta_k\,, \\
M_2(\theta) &= \sum_{k=1}^{K} \omega_k E[g''\big(\langle X, \mu_k\rangle + b_k\big)]\ \beta_k \otimes \beta_k\,, \\
M_3(\theta) &= \sum_{k=1}^{K} \omega_k E[g^{(3)}\big(\langle X, \beta_k\rangle + b_k\big)]\ \beta_k \otimes \beta_k \otimes \beta_k\,.
\end{aligned}$$

Voir l'article pour les détails des calculs. Il est alors naturel d'utiliser un estimateur des moindres carrés, minimisant les écart des $M_i(\theta)$ aux $\widehat{M_i}$.

2.2 Précisions

Considérant la somme de carrés suivante :

$$Q_n(\theta) = \sum_{j\in[d]} \Big\{ \widehat{M_1}[j] - M_1(\theta)[j] \Big\}^2 + \sum_{j,k\in[d]} \Big\{ \widehat{M_2}[j,k] - M_2(\theta)[j,k] \Big\}^2 + \sum_{j,k,l\in[d]} \Big\{ \widehat{M_3}[j,k,l] - M_3(\theta)[j,k,l] \Big\}^2,$$

avec $[d] = [1, \ldots, d]$ l'ensemble des entiers de 1 à d, d étant la dimension des covariables. Une première idée consiste à minimiser $Q_n(\theta)$ directement :

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}}\ Q_n(\theta). \tag{1}$$

Les termes dans la somme contribuent inégalement, car il y a plus de combinaisons d'indices pour les moments d'ordre supérieur. On peut alors penser à pondérer chaque groupe de termes : cela mène déjà à certaines améliorations. Cependant, on peut aller plus loin en réécrivant le problème de minimisation. Définissons pour tout θ et $i = 1, \ldots, n$

$$\tilde{M}_i(\theta) = Y_i \begin{pmatrix} X_i \\ X_i \otimes X_i - \sum_{j\in[d]} e_j \otimes e_j \\ X_i \otimes X_i \otimes X_i - \sum_{j\in[d]} \big[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i \big] \end{pmatrix} - \begin{pmatrix} M_1(\theta) \\ M_2(\theta) \\ M_3(\theta) \end{pmatrix}$$

comme un vecteur colonne. Posons alors le problème de minimisation suivant (Hansen 1982).

Soit W une matrice symétrique définie positive de taille $d + d^2 + d^3$. Écrivons

$$Q_n^W(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} {}^t\tilde{M}_i(\theta) \right) W \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{M}_i(\theta) \right),$$

et définissons

$$\widehat{\theta}_n^W = \underset{\theta \in \Theta}{\operatorname{argmin}}\ Q_n^W(\theta). \tag{2}$$

On remarque alors que si W est la matrice identité, on retrouve $Q_n(\theta)$. Si W est diagonale on obtient la première idée mentionnée ci-dessus.

3 Algorithme

À ce stade se pose la question du choix de la matrice W. Hansen (1982) démontre que la matrice W minimisant la variance asymptotique de l'estimateur est donnée par $W(\theta^\star) = (\mathrm{E}[\tilde{M}_i(\theta)^t\tilde{M}_i(\theta)])^{-1}$. Cette matrice optimale n'est pas accessible mais peut être approchée empiriquement par

$$\hat{W}(\hat{\theta}) = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{M}_i(\hat{\theta})^t\tilde{M}_i(\hat{\theta}) \right)^{-1},$$

avec $\hat{\theta}$ une estimation préliminaire des paramètres, par exemple avec $W = Id$.

L'algorithme consiste alors à obtenir une première estimation des paramètres (avec $W = Id$ par exemple), elle-même initialisée avec les directions de la matrice β estimées depuis les données : $\mu_k = \beta_k / \|\beta_k\|$. On recalcule alors la matrice W selon la formule précédente, puis survient la seconde et dernière itération.

Algorithme d'estimation des paramètres
Entrée: X, Y, K, g
1 : Estimer les directions $\mu_1, \ldots, \mu_K$ via l'algorithme $InitDir$
2 : Minimiser $Q_n^{W_{\texttt{init}}}(\theta)$ en partant partant des directions trouvées en 1.
(S'arrêter ici si $W_{\texttt{init}}$ est considérée assez précise)
3 : Re-calculer W en utilisant les paramètres ω, β et b obtenus
4 : Ré-exécuter l'étape 2.
Sortie: Les paramètres estimés $\widehat{\theta}$

Voir l'article Auder et al (2020) concernant les détails de l'étape d'estimation des directions (un calcul algébrique qui sort du cadre de cette communication).

4 Expériences numériques

Nous comparons l'algorithme du paragraphe précédent avec celui plus classique basé sur la maximisation de la log-vraisemblance implémenté dans le package R flexmix (2019). Des jeux de données simulés sont utilisés à cet effet, de dimension 5 et 10. En dimension 5 nous simulons 2 groupes, puis 3 en dimension 10. Les données sont prises équiréparties dans chaque groupe, et les matrices β contiennent des coefficients aléatoires variant entre -4 et 4. Par exemple dans le cas $d = 10$:

$$\beta = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & 1 \\ -1 & 0 & 3 \\ 0 & 1 & -1 \\ 3 & 0 & 0 \\ 4 & -1 & 0 \\ -1 & -4 & 2 \\ -3 & 3 & 0 \\ 0 & 2 & 1 \\ 2 & 0 & -2 \end{pmatrix}$$

Afin de comparer les erreurs d'estimation, nous affichons dans la table suivante l'erreur L^1 sommée sur les composantes de chaque paramètre. C'est-à-dire que pour ω l'erreur affichée est $\sum_{k=1}^{K} |\omega_k - \hat{\omega}_k|$. De même pour b, et pour β en considérant cette matrice colonne par colonne. Les paramètres sont obtenus en moyennant sur $N = 100$ réplications pour différentes valeurs de n. Ensuite, on retire 2% des plus grandes valeurs, qui sont très rares et biaiseraient trop le résultat visuel − pour les deux méthodes.

Les performances sont comparables, avec un léger avantage à flexmix dans le cas logit, et à notre algorithme − package R morpheus (2020) − dans le cas probit (non montré ici). Les temps d'exécution sont en général meilleurs pour morpheus. Étant donné que l'estimation par maximum de vraisemblance est asymptotiquement optimale, c'est un résultat très encourageant.

	$d=5$				$d=10$				
n	p	β_1	β_2	b	p	β_1	β_2	β_3	b
5.10^3	1.8e-2	2.1e+0	1.6e+0	3.9e-1	4.9e-2	5.2e+0	3.8e+0	1.2e+1	7.8e+0
	6.0e-4	2.7e-1	7.0e-2	6.3e-3	1.3e-1	1.6e+2	1.8e+2	1.4e+0	4.3e+0
10^4	2.5e-2	7.1e-1	9.7e-1	4.6e-1	3.3e-2	3.9e+0	3.4e+0	7.4e+0	2.7e+0
	1.8e-3	3.3e-2	4.2e-2	6.9e-3	1.3e-1	2.2e+0	1.8e+0	4.8e-1	7.3e-2
10^5	5.8e-2	4.1e-1	2.6e-1	3.5e-2	1.7e-2	1.1e+0	6.8e-1	2.0e+0	2.7e-1
	5.9e-5	2.2e-2	1.7e-2	2.4e-3	1.3e-1	4.6e-2	9.6e-2	9.3e-2	5.5e-3
5.10^5	1.9e-2	2.0e-1	8.0e-2	9.4e-3	1.6e-2	9.1e-1	1.0e+0	7.9e-1	7.6e-2
	2.3e-4	5.0e-3	6.1e-3	3.9e-3	1.3e-1	2.8e-2	1.7e-2	1.7e-2	2.5e-3
10^6	7.0e-2	5.3e-1	4.0e-1	1.9e-2	7.5e-3	7.8e-1	8.7e-1	2.7e-1	7.8e-2
	7.0e-5	2.5e-3	7.0e-3	2.5e-3	1.3e-1	5.4e-2	2.0e-2	1.0e-2	3.4e-3

Table 1: Lien *logit*. Somme des erreurs pour notre algorithme (en haut) et flexmix (en bas) moyennées sur $N = 100$ réplications, pour des valeurs croissantes de n.

Bibliographie

A. Anandkumar et al. (2014), Tensor decompositions for learning latent variables models, *Journal of machine learning*, 15, 2773–2832.

B. Auder et al. (2020), Least squares moment identification of binary regression mixtures models, *Arxiv: https://arxiv.org/abs/1811.01714v2*, submitted.

L. P. Hansen (1982), Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029–1054.

B. Gruen et al. (2019), flexmix: Flexible Mixture Modeling, *CRAN: https://CRAN.R-project.org/package=flexmix*.

B. Auder et M-A. Loum (2020), morpheus: Estimate Parameters of Mixtures of Logistic Regressions, *CRAN: https://CRAN.R-project.org/package=morpheus*.

Échantillonage Préférentiel Adaptatif Multiple Tempéré pour l'analyse des Distributions Spectrales D'énergie des Galaxies

Grégoire Aufort [1,2] & Pierre Pudlo [1] & Denis Burgarella [2]

[1] *Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*
[2] *Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France*
gregoire.aufort@univ-amu.fr / pierre.pudlo@univ-amu.fr / denis.burgarella@lam.fr

Résumé. Nous étudions le problème de l'inférence bayésienne pour l'analyse des Distributions Spectrales d'énergie des galaxies. Nous proposons un modèle statistique et une nouvelle version de lÉchantillonage Préférentiel Adaptatif Multiple qui stabilise la convergence de l'algorithme en introduisant une suite de distributions cibles auxiliaires bien choisies. Cette modification permet à la fois une diminution du nombre nécessaires d'évaluations de la vraisemblance ainsi qu'une plus grande robustesse à une mauvaise initialisation et au fléau de la dimension. L'association de cet algorithme et d'une approximation de la vraisemblance à l'aide de réseaux de neurones permet d'accélérer et d'automatiser l'analyse.

Mots-clés. Inférence Bayésienne,échantillonage Préférentiel Adaptatif, Tempering, Astrophysique des galaxies

We study bayesian inference for galaxy Spectral Energy Distribution analysis. We propose a statistical model and a new version of the Adaptive Importance Sampling algorithm stabilising convergence by introducing a well chose sequence of auxiliary target distributions. This modification allows for reducing the number of likelihood evaluations, moderates the curse of dimensionality and is more robust against poor initializations. The combination of this algorithm with a Neural Network approximation of the likelihood allows to accelerate and automate inference.

Keywords. Bayesian inference, Adaptive Importance Sampling, Tempering, Astrophysics of galaxies

L'étude de la distribution spectrale d'énergie (Spectral Enegy Distribution ou SED) d'une galaxie est le meilleur moyen d'étude de ses propriétés physiques. Des modèles physiques complexes permettent en effet de relier les principales propriétés d'histoire et de composition des éléments constitutifs d'une galaxie (étoiles, gaz, poussières) à sa SED. On observe ensuite partiellement la SED via differents instruments de mesures (télescopes, spectrographes) à différentes longueurs d'ondes, et on cherche à estimer les paramètres des modèles physiques à partir de ces observations partielles.
Du point de vue statistique, il s'agit donc pour chaque galaxie observée, d'ajuster une courbe théorique, dépendant de paramètres physiques θ, à des mesures échantillonnées

irrégulièrement le long de cette courbe, et issues de deux procédés de mesure, la spectrographie et la photométrie. Nous proposons une méthode d'inférence bayésienne basée sur *(i)* l'approximation par un réseau de neurones des courbes théoriques issues du modèle physique, donc le coût de calcul était trop long, sur *(ii)* un modèle de bruit gaussien autour de cette courbe théorique, et sur *(iii)* une pondération des différents points de mesure sur la SED, pour tenir compte des deux moyens de mesures, et de l'hétérogénéité de l'échantillonnage en longueur d'onde le long de la courbe.

Cette configuration nécessite l'emploi d'outils d'intégration numérique adaptés. Le parallèlisme naturel des réseaux de neurones sur GPU n'est absolument pas exploité par un algorithme MCMC, qui doit attendre à chaque itération la nouvelle valeur proposée de θ pour lancer le calcul de la courbe théorique. En revanche, les algorithmes particulaires comme Population Monte Carlo (Cappé *et al.* 2004), Sequential Monte Carlo (Doucet *et al.*(2001)) ou Adaptive Multiple Importance Sampling (Cornuet *et al.* 2012) sont adaptés ici, puisqu'ils envoient, à chaque itération, toute une population de valeurs de θ pour lesquelles ils souhaitent connaître la valeur de la SED théorique. Nous nous concentrons ici sur les algorithmes échantillonages préferentiels adaptatifs, mais l'initialisation et la calibration de la loi de proposition de ces algorithmes est un problème difficile. Notre contrainte est de proposer un algorithme suffisamment stable pour pouvoir traiter automatiquement beaucoup de galaxies et obtenir autant de distribution a posteriori que de galaxies à étudier. Par exemple, l'algorithme d'initialisation proposé par Cornuet *et al.* (2012) dans la section 4 est instable, et très coûteux en nombre d'évaluations de la SED théorique. Nous proposons donc une nouvelle méthode qui ralentit la course vers la loi a posteriori et stabilise la calibration de la loi de proposition.

La première partie de ce résumé présente le modèle statistique, et la seconde partie présente notre algorithme d'échantillonage préférentiel.

1 Modélisation de la distribution spectrale d'énergie d'une galaxie

Nous décrivons ici succinctement les données, le calcul de la courbe théorique que l'on ajuste aux données et le modèle statistique.

Pour une galaxie d'étude, on dispose d'un vecteur $x \in \mathbb{R}^d$ de flux d'énergie mesurés à différentes longueurs d'onde. Ce vecteur $x = (x_{\text{ph}}, x_{\text{sp}}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ se décompose en deux sous-vecteurs qui correspondent aux deux procédés de mesure (photométrie, spectrométrie) pour un flux d'énergie. La Figure 1 montre la répartition des longueurs d'onde observées par chaque procédé de mesure et des courbes théoriques ajustées aux données. Pour une valeur θ du vecteur des paramètres physiques, on note $\text{SED}(\theta) \in \mathbb{R}^d$ les valeurs de flux proposées par le modèle physique aux longueurs d'onde observées, voir Figure 2. Et,

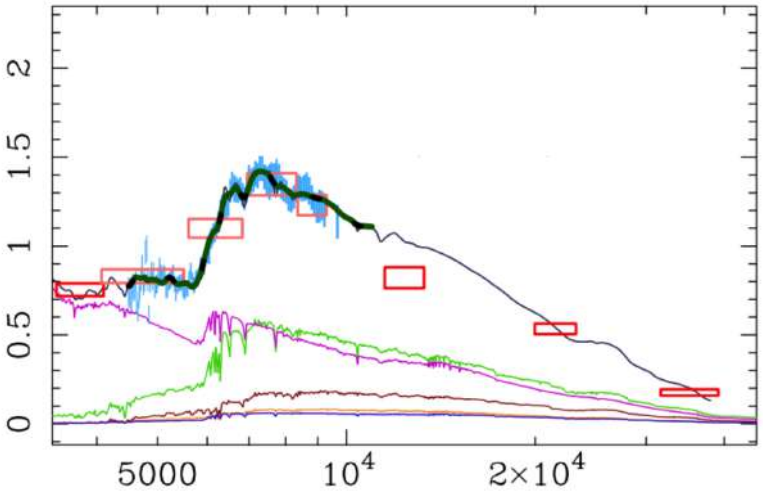


Figure 1: **Illustration des différents types de mesures**.
Les rectangles rouges représentent des mesures photométriques (peu nombreuses mais couvrant une part importante du spectre), les bleus des mesures spectroscopiques (beaucoup plus nombreux mais beaucoup plus concentrés). La courbe noire est une SED théorique ajustée aux données par maximum de vraisemblance, et les courbes colorées représente la décomposition de la SED selon ses contributions (étoiles, poussière, gaz). L'une des mesures photométrique n'est compatible avec aucune SED théorique

comme pour les observations, on décompose $\mathrm{SED}(\theta) = (\mathrm{SED}_{\mathrm{ph}}(\theta), \mathrm{SED}_{\mathrm{sp}}(\theta)) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.

Le modèle de bruit dépend du procédé de mesure. Ce qui revient à considérer deux vraisemblances partielles

$$f_{\mathrm{ph}}\big(x_{\mathrm{ph}}\big|\theta\big) = \mathcal{N}_{d_1}\big(x_{\mathrm{ph}}\big|\mathrm{SED}_{\mathrm{ph}}(\theta), \Sigma_{\mathrm{ph}}\big),$$
$$f_{\mathrm{sp}}\big(x_{\mathrm{sp}}\big|\theta\big) = \mathcal{N}_{d_2}\big(x_{\mathrm{sp}}\big|\mathrm{SED}_{\mathrm{sp}}(\theta), \Sigma_{\mathrm{sp}}\big).$$

Nous supposons que ces deux erreurs gaussiennes multivariées sont indépendantes, que la matrice Σ_{ph} est diagonale, alors que nous considérons une structure de corrélation dans Σ_{sp} qui tient compte de la distance courte entre les différentes longueurs d'onde observées par spectroscopie. Par ailleurs, les observations issues de la spectroscopie sont beaucoup plus nombreuses ($d_2 \gg d_1$), et concentrées dans une petite partie du spectre. Pour ne pas sur-ajuster la courbe théorique dans cette région très riche en données, nous proposons de pondérer les contributions à la vraisemblance de chacun des procédés de mesure via leurs taux de couverture relatifs, $\Delta\lambda_{\mathrm{ph}}$ et $\Delta\lambda_{\mathrm{sp}}$. En posant $r = \Delta\lambda_{\mathrm{sp}}/\Delta\lambda_{\mathrm{ph}}$, la vraisemblance que nous considérons est donc

$$L(\theta|x) = f_{\mathrm{ph}}\big(x_{\mathrm{ph}}\big|\mathrm{SED}_{\mathrm{ph}}(\theta)\big)\, f_{\mathrm{sp}}^{r}\big(x_{\mathrm{sp}}\big|\mathrm{SED}_{\mathrm{sp}}(\theta)\big), \tag{1}$$

ce qui revient à augmenter artificiellement la matrice de covariance Σ_{sp}. Enfin, nous considérons une distribution a priori $p(\theta)$ généralement uniforme pour tenir compte de contraintes physiques sur les valeurs des paramètres (positivité, âge,...).

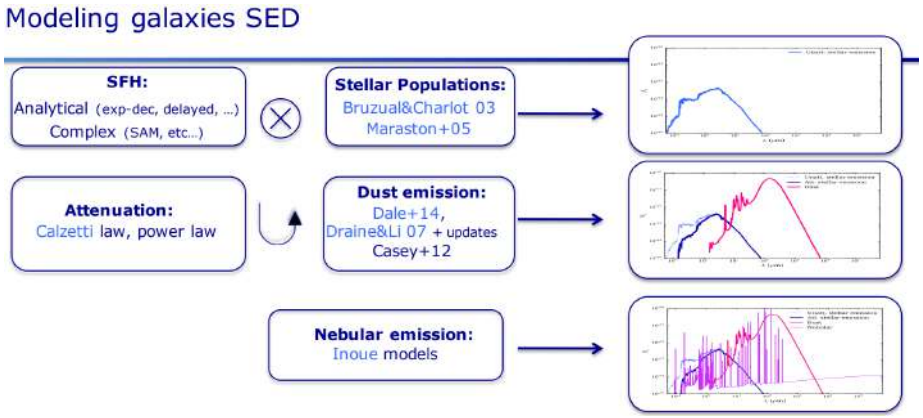


Figure 2: **Calcul de la SED théorique**.
Le logiciel CIGALE (Boquien *et al*, 2019 ; où les références de la figure sont données) calcule une distribution spectrale d'énergie (SED) en trois étapes (les trois lignes). La première étape convolue l'histoire de formation stellaire (SFH) par un modèle de population stellaire pour obtenir la courbe théorique bleue. La seconde étape considère les effets de la poussière interstellaire sur la lumière ainsi émise : cela donne une courbe de flux (en rouge). La troisième étape calcule les emissions dûes au gaz nébulaire. La courbe théorique résultante est la somme de ces trois composantes.

2 Échantillonage Préférentiel Adaptatif Multiple Tempéré

Décrivons maintenant l'algorithme d'échantillonnage de la loi a posteriori $\pi(\theta) \propto p(\theta)L(\theta|x)$.

Rappelons que l'évaluation numérique de la vraisemblance repose sur le calcul de SED(θ). Ce dernier vecteur, issu d'une fonction boîte noire, est en fait approché par un réseau de neurones, et reste coûteux à calculer. Pour réduire le coût de recours au réseau (via le GPU), on évalue cette fonction par lots de valeurs de θ. Nous nous appuyons donc sur un algorithme d'échantillonnage préférentiel qui échantillonne l'espace des paramètres à l'aide d'une loi de proposition $q(\theta)$ et pondère les tirages par $w(\theta) = \pi(\theta)/q(\theta)$.

Le choix de la loi q est crucial pour la qualité de l'approximation, la meilleure loi étant la loi a posteriori elle-même. Différents algorithmes itératifs, qui améliorent progressivement la distribution d'échantillonnage q_t de l'itération t, ont été proposés. (voir par exemple Koblents et Joaquin (2012), Bugallo *et al.* (2017), Marin *et al.* (2019)) À chaque itération, nous calibrons une loi de mélange pour définir q_{t+1}, comme dans Cappé *et al.*(2008) et Douc *et al.*(2007). La stratégie classique de la littérature est de choisir q_{t+1} qui minimise la divergence de Kullback-Leibler $D_{KL}(\pi||q_{t+1})$, en approchant la loi cible $\pi(\theta)$ par l'échantillon pondéré issu de la précédente itération. Malheureusement, cette méthode ne résout que partiellement le problème puisqu'elle repose sur le choix de la première loi q_0, et peut se révéler numériquement instable.

Nous proposons de ralentir la course vers la loi cible $\pi(\theta)$ de deux façons : *(i)* à la

fin de l'itération t, nous calibrons la loi q_{t+1} en minimisant sa divergence $D_{KL}(\pi_\beta||q_{t+1})$ comme dans Cappé *et al.*(2008) mais avec une version tempérée de la cible $\pi_\beta(\theta)$ et *(ii)*, nous corrigeons cette loi d'échantillonnage pour que les particules θ associées à des petits poids $w_{t+1}(\theta) = \pi(\theta)/q_{t+1}(\theta)$ puissent être utilisées lors de la calibration.

Ainsi, pour calibrer q_{t+1}, nous pourrions chercher la loi de mélange gaussien $\tilde{q}_{t+1}$ qui minimise une version empirique de $D_{KL}(\pi_\beta||q_{t+1})$, où

$$\pi_\beta(\theta) = \pi(\theta)^\beta q_t(\theta)^{1-\beta} \tag{2}$$

et où $\beta \in [0;1]$ est auto-calibré. Cette pseudo-cible tempérée est un compromis entre la loi a posteriori (cible finale) et la distribution auxiliaire utilisée à l'étape précédente. En effet, la divergence $D_{KL}(\pi||\pi_\beta)$ de la cible contre la pseudo-cible tempérée (2) est une fonction décroissante de β. L'inverse température β est auto-calibrée à chaque itération pour assurer la qualité de l'approximation de π_β par q_{t+1} via le critère d'Effective Sample Size (ESS). Mais de nombreux poids $w_t(\theta) = \pi(\theta)/q_t(\theta)$ qui apparaissent dans la version empirique de $D_{KL}(\pi_\beta||q_{t+1})$ restent très faibles pour contribuer significativement à l'approximation.

Donc, nous remplaçons tous les poids $w_t^\beta(\theta) = \pi_\beta(\theta)/q_t(\theta)$ par

$$\widetilde{w}_t^\beta(\theta) = \max\left(s, w_t^\beta(\theta)\right), \quad \text{où}$$

s est le quantile d'ordre 60% des poids $w_t^\beta(\theta)$ calculés sur les tirages de l'itération t. Cela revient à remplacer la loi π_β par une autre pseudo-cible $\tilde{\pi}_\beta$ donnée par

$$\tilde{\pi}_\beta(\theta) \propto s\, q_t(\theta)\mathbf{1}\{\theta \in E\} + \pi_\beta(\theta)\mathbf{1}\{\theta \notin E\}, \quad \text{avec } E = \left\{\theta \ : \ \frac{\pi_\beta(\theta)}{q_t(\theta)} < s\right\}. \tag{3}$$

Et à chercher la loi de mélange q_{t+1} qui minimise une version empirique de $D_{KL}(\tilde{\pi}_\beta||q_{t+1})$. Il s'agit d'un second frein à la course vers la cible $\pi(\theta)$ puisque les particules ainsi pondérées visent une pseudo-cible $\tilde{\pi}_\beta$, mais qui avance dans la bonne direction puisque, *(i)* q_{t+1} est une approximation par une loi de mélange de $\tilde{\pi}_\beta$ et *(ii)*, si $s \leq 1$,

$$D_{KL}(\pi_\beta||\tilde{\pi}_\beta) \leq D_{KL}(\pi_\beta||q_t).$$

Au final, la mise en place de ces deux freins à la course vers la cible $\pi(\theta)$ prévient l'échec de l'étape de mise à jour de l'algorithme itératif en empêchant la dégénérescende des poids, tout en gardant faible le nombre total de tirage de θ, donc d'évaluations coûteuses de la vraisemblance $L(\theta|x)$. Nous montrons sur des simulations que cela permet de mettre en œuvre une procédure d'échantillonnage préférentiel, même en dimension 50. De façon plus surprenante, nos simulations montreront aussi une certaine robustesse quant au choix de la loi auxiliaire initiale q_0 (voir Figure 3).

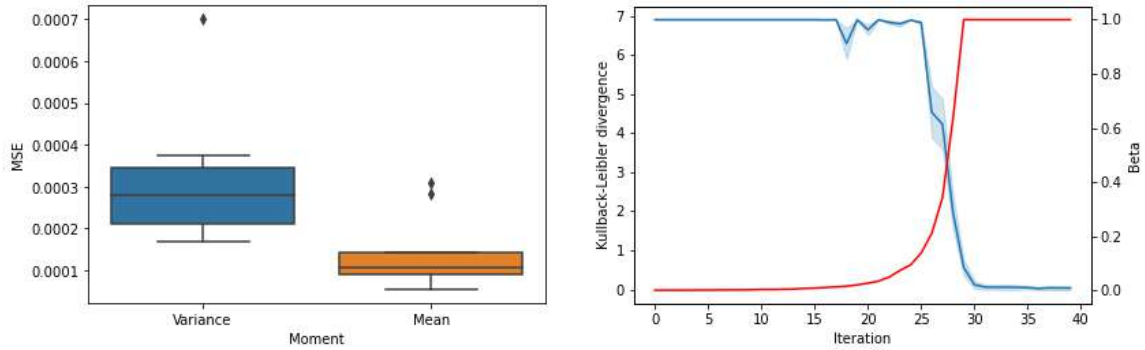


Figure 3: **Un exemple jouet**.
Résultats de notre algorithme pour l'estimation de la moyenne et de la variance d'une gaussienne en dimension 20 $\pi = \mathcal{N}(50_{20}, I_{20})$, avec une très mauvaise initialisation $\mathcal{N}(0_{20}, 200 \times I_{20})$, avec un total de seulement $40,000$ tirages (40 itérations de 1000 tirages chacune). À gauche l'erreur quadratique moyenne de l'estimation de la variance et de la moyenne (10 répétitions de l'expérience). À droite lévolution de $D_{KL}(\pi||q)$ (en bleu) et de β_t (en rouge) au cours des itérations durant l'une des répétitions.

Bibliographie

Boquien, M. Burgarella, D. Roehlly, Y. Buat, V. Ciesla, L. Corre, D. Inoue, A. K. et Salas, H. (2019).,CIGALE: a python Code Investigating GALaxy Emission,*Astronomy & Astrophysics*, 622, A102

Bugallo, M., Elvira, V., Martino, L., Luengo, D., Miguez, J. et Djuric, P. (2017). Adaptive Importance Sampling: The past, the present, and the future. *IEEE Signal Processing Magazine.*

Cappé, O., Guillin, A., Marin, J.-M., et Robert, C. (2004). Population Monte Carlo.*Journal of Computational and Graphical Statistics*, 13:907-929

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., et Robert, C. (2008). Adaptive importance sampling in general mixture classes.*Statistics and Computing*, 18:587-600.

Ciesla, L. Elbaz, D. et Fensch, J. (2017) The SFR-M main sequence archetypal star-formation history and analytical models. *Astronomy & Astrophysics* 608, A41.

Cornuet, J., Marin, J.-M., Mira, A., et Robert, C. (2012). Adaptive Multiple Importance Sampling. *Scandinavian Journal of Statistics*, 39(4), 798-812.

Douc, R., Guillin, A., Marin, J.-M. et Robert, C. P. (2007) Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics.* 35 (1) 420 - 448,

Doucet, A. , de Freitas, N. et Gordon, N. (2001) Sequential Monte Carlo Methods in Practice. *Springer.*

Koblents, E. et Miguez, J. (2012). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing.*

Marin,J.-M., Pudlo, P., Sedki, M. (2019) Consistency of the Adaptive Multiple Importance Sampling. *Bernoulli*

Régression fonctionnelle linéaire et non paramétrique basée sur des projections orthogonales aléatoires

Bilel Bousselmi [1], Jean-François Dupuy [2] and Abderrazek Karoui [3]

[1] *Université de Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France. (Bilel.bousselmi@insa-rennes.fr)*

[2] *Université de Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France. (Jean-Francois.Dupuy@insa-rennes.fr)*

[3] *Université de Carthage, Département de Mathematiques, Faculté des Sciences de Bizerte, Tunisie. (Abderrazek.Karoui@fsb.rnu.tn)*

Résumé. Ce travail a un double objectif. Dans un premier temps, nous développons un schéma basé sur les polynômes orthogonaux de Jacobi pour la résolution stable du problème de régression non paramétrique. Dans un deuxième temps, nous développons un estimateur basé sur une projection aléatoire orthogonale pour la résolution stable du problème de régression fonctionnelle linéaire (LFR). Plus précisément, le premier schéma nous fournit une famille à deux paramètres d'estimateurs de régression non paramétrique. Ces estimateurs sont construits en utilisant le noyau de Christoffel issu de la famille de polynôme de Jacobi $\{P_k^{(\alpha,\beta)}, 0 \leq k \leq N\}$ où $\alpha, \beta \geq -\frac{1}{2}$. Une analyse de convergence de cet estimateur à deux paramètres est faite sous l'hypothèse que la vraie fonction de régression a une certaine régularité de Lipschitz ou qu'il s'agit de la restriction à $I = [-1, 1]$ d'une fonction à bande limitée. De plus, nous montrons que notre deuxième estimateur LFR basé sur les projections aléatoires orthogonales est stable. En effet, on montrera que la matrice de projection aléatoire associée est bien conditionnée. Enfin, nous illustrons les différents résultats de ce travail par quelques simulations numériques.

Mots-clés: Régression non paramétrique, projection aléatoire, noyau de Jacobi, régression fonctionnelle linéaire, pseudo-inverse aléatoire, nombre de conditionnement.

Abstract. The aim of this work is twofold. Firstly, we develop a Jacobi polynomials based scheme for the stable solution of non parametric regression problems. Secondly, we develop an orthogonal projection based estimator of the stable solution of linear functional regression (LFR) problem. More precisely, the first scheme provides us with a two parameters family of non parametric regression estimators. These estimators are constructed by using a random projection kernel associated with a finite orthonormal set of Jacobi polynomials $\{P_k^{(\alpha,\beta)}, 0 \leq k \leq N\}$ where $\alpha, \beta \geq -\frac{1}{2}$. A convergence analysis of these two parameters estimators is done under the assumption that the true regression function has some Lipschitz regularity or is the restriction to $I = [-1, 1]$ of bandlimited function. Moreover, we show that our random orthogonal projection based LFR estimator is stable. This is done by showing that the associated random projection matrix is well conditioned. Finally, we illustrate the different results of this work by some numerical simulations.

Keywords: Nonparametric regression, orthogonal projection, Jacobi kernel, linear functional regression, random pseudo-inverse, condition number.

In the first part of this work, we consider the following nonparametric regression model:

$$Y_i = f(X_i) + \eta_i, \quad 1 \leq i \leq n,$$

where $(X_i)_{1\leq i\leq n}$ are random variables (or inputs) and the noise terms $(\eta_i)_{1\leq i\leq n}$ are i.i.d. real-valued centred random variables. For simplicity, we will assume that the X_i take values in the interval $I = [-1, 1]$. Nonetheless, we show how our nonparametric regression estimator can be adapted to handle the multivariate case where $X_i \in \mathbb{R}^d$, $d \geq 2$. For a given complete metric space $\mathcal{X}$, an output space $\mathcal{Y}$, by

following the standard notations (see for example [8]) and letting ρ_X denote the marginal probability measure over $\mathcal{X}$, the true regression function associated with the previous regression problem is given by

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X},$$

where $d\rho(y|x)$ is the conditional distribution of Y given $X = x$.
In the first part of this work, our aim is to develop a kernel projection based scheme that provides convenient and stable estimates of the regression function f, provided that this later satisfies some smoothness property. For two real parameters $\alpha, \beta > -1/2$ and a positive integer N, the Jacobi polynomials kernel is given by:

$$K_N^{\alpha,\beta}(x,y) = \sum_{k=0}^{N} \left[\widetilde{P}_k^{(\alpha,\beta)}(X_i)\widetilde{P}_k^{(\alpha,\beta)}(x)\right] = A_N^{(\alpha,\beta)} \begin{cases} \dfrac{\widetilde{P}_{N+1}^{(\alpha,\beta)}(x)\widetilde{P}_N^{(\alpha,\beta)}(y) - \widetilde{P}_N^{(\alpha,\beta)}(x)\widetilde{P}_{N+1}^{(\alpha,\beta)}(y)}{x-y}, & x \neq y \\ \widetilde{P}_{N+1}^{(\alpha,\beta)}(x)'\widetilde{P}_N^{\alpha}(x) - \widetilde{P}_N^{(\alpha,\beta)}(x)'\widetilde{P}_{N+1}^{(\alpha,\beta)}(x), & x = y \end{cases},$$

where $x, y \in I$, $A_N^{(\alpha,\beta)} = \dfrac{2}{2N+\alpha+\beta+2}\sqrt{\dfrac{(N+1)(N+\alpha+\beta+1)(N+\alpha+1)(N+\beta+1)}{(2N+\alpha+\beta+1)(2N+\alpha+\beta+3)}}$, and the $\widetilde{P}_k^{(\alpha,\beta)}$ are the usual orthonormal Jacobi polynomials with parameters $\alpha, \beta > -\frac{1}{2}$ and degree $k \geq 0$.

By using a random training set $\{(X_i, Y_i),\ 1 \leq i \leq n\}$, we construct an empirical projection operator associated with the Jacobi kernel. More precisely, we consider the weight function $\omega_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$ and the associated weighted $L^2_\omega(I)-$space of real valued, measurable functions f and satisfying $\|f\|^2_\omega = \displaystyle\int_I |f(x)|^2\omega_{\alpha,\beta}(x)\,dx < +\infty$. For a positive integer $n \in \mathbb{N}$, let $\{X_i,\ 1 \leq i \leq n\}$ be a sampling set of i.i.d random variables following the beta distribution $B(\alpha+1, \beta+1)$ over I. Then, we define our estimator of the regression function f, based on Jacobi kernel by:

$$\widehat{f}_{n,N}^{\alpha,\beta}(x) \quad = \quad \frac{2^{\alpha+\beta+1}B(\alpha+1,\beta+1)}{n}\sum_{i=1}^{n} Y_i K_N^{\alpha,\beta}(X_i, x), \quad x \in I.$$

To give an error estimate of this estimator, we need to define the following constants. For $\alpha, \beta \geq -\frac{1}{2}$ let $c_{\alpha,\beta} = \dfrac{\alpha+\beta+1}{2}$, $\eta_{\alpha,\beta} = \dfrac{\exp\left(\frac{2\max(\mu,0)}{12} + \frac{\max(\mu^2+\alpha\beta,0)}{8}\right)}{2^{(\alpha+\beta)/2}\Gamma(\mu+1)}$, $\boldsymbol{\eta}_n = \max_i |\eta_i|$, $\mu = \max(\alpha,\beta)$, $\tau = \min(\alpha,\beta)$. $M_{f,N}^{\alpha,\beta}(\boldsymbol{\eta}_n) = \eta_{\alpha,\beta}2^{(\alpha+\beta+1)}\sqrt{\frac{1+c_{\alpha,\beta}}{2\mu+1}}\Big(\beta(\alpha+1,\beta+1)\Big)(N+1)^{\mu+1}(\|f\|_\infty + \boldsymbol{\eta}_n) + \sqrt{2}\|f\|_{\omega_{\alpha,\beta}}$. The following theorem provides us with an estimate for the squared L^2-error of our estimator $\widehat{f}_{N,n}^{\alpha,\beta}(\cdot)$.

Theorem 1. *Under the previous notations and assumptions, assume that f has p continuous derivatives on I and its $p-th$ derivative $f^{(p)} \in Lip(\gamma)$, with $p+\gamma \geq \max(\mu+\frac{1}{2}, \frac{1}{2}-\tau)$. Then for any $0 < \delta < 1$ and for sufficiently large values of n, N, we have with probability at least $1-\delta$,*

$$\|f - \widehat{f}_{n,N}^{\alpha,\beta}\|_\omega \lesssim \frac{\log N}{n^{p+\gamma}} + \frac{M_{f,N}^{\alpha,\beta}(\boldsymbol{\eta}_n)}{\sqrt{n}}\sqrt{\log\left(\frac{2}{\delta}\right)}. \tag{1}$$

Next, we assume that for some $c > 0$, f is the restriction to I of a $c-$bandlimited function $\widetilde{f}$, that is $\widetilde{f}$ belongs to the Paley-Wiener space $\mathcal{B}_c$, defined as the set of functions of $L^2(\mathbb{R})$ with Fourier transforms supported on the interval $[-c, c]$. Then, for $\beta = \alpha \geq -\frac{1}{2}$ and any $0 < \delta < 1$, we have with probability at least $(1-\delta)^2$:

$$\|f - \widehat{f}_{n,N}^{\alpha,\alpha}\|_{L^2(I)} \leq \frac{M_{f,N}^{\alpha,\alpha}(\boldsymbol{\eta}_n)}{\sqrt{n}}\sqrt{\log\left(\frac{2}{\delta}\right)} + \frac{\gamma(\alpha)}{\sqrt{c}}\left(1 + \frac{1}{2\ln\left(\frac{2N+4}{ec}\right)}\right)\left(\frac{ec}{2N+2}\right)^{N+2}\left\|\widetilde{f}\right\|_{L^2(\mathbb{R})}, \tag{2}$$

where $\gamma(0)=1, \gamma(\alpha)=2^{-\frac{3}{2}\alpha+\frac{1}{4}}e^{-\alpha-\frac{1}{4}}$, $\alpha\neq 0$. (see [2] for more details)

Remark 1. *The quantity* $\boldsymbol{\eta}_n=\max\limits_{1\leq i\leq n}|\eta_i|$, *given in* (1) *and* (2) *depends on* n. *Nonetheless, in practice and independently of* n, *the noises* η_i *are uniformly bounded with high probability. This is the case for example when the* η_i *are i.i.d. copies of the largely used centered Gaussian noise model with variance* σ^2. *In this case, for any fixed* $k_0>0$, *we have* $|\eta_i|\leq k_0\sigma$ *with probability at least* $1-erf\Big(\dfrac{k_0}{\sqrt{2}}\Big)\approx 1-\dfrac{e^{-k_0^2/2}}{k_0\sqrt{\pi/2}}$ *which is a very high probability even for reasonable small values of* k_0.

Also, to overcome the problem of handling random sampling set, associated with an unknown marginal distribution law ρ_X, one may use the well known transformation technique of random sampling laws. This technique is briefly described as follows. We first assume that random sampling points X_i are i.i.d. copies of a random variable with known cumulative distribution function (CDF) $F_X(\cdot)$. Then it is easy to see that the $F_X(X_i)$ follow the uniform law over $(0,1)$. Since the CDF of the Beta distribution $B(\alpha,\alpha)$ is given by the regularized incomplete Beta function $I_x(\alpha,\alpha)$, and this later is invertible, then the transformed sampling points

$$Z_i=I_x^{-1}(\alpha+1,\beta+1)\big(F_X(X_i)\big),\quad 1\leq i\leq n$$

follow the $Beta(\alpha+1,\beta+1)-$distribution. For the case of an unknown sampling law ρ_X, it suffices to replace the true CDF function $F_X(\cdot)$ by an accurate estimate, in the previous equality.

Finally, we note that our Jacobi polynomials kernel projection estimator $\widehat{f}_{n,N}^{\alpha,\beta}$ which was developed and studied in the univariate case can be straightforwardly generalized to the multivariate case, where the random sampling sets $\{X_i, 1\leq i\leq n\}\subset\mathbb{R}^d$. In fact, it suffices to replace each Jacobi polynomial $\widetilde{P}_k^{(\alpha,\beta)}$ by its tensor product $d-$dimensional version

$$\Phi_{\boldsymbol{m}}^{(\alpha,\beta)}(\boldsymbol{x})=\prod_{j=1}^{d}\widetilde{P}_{\boldsymbol{m}}^{(\alpha,\beta)}(x_j),\quad \boldsymbol{x}=(x_1,\ldots,x_d)\in I^d,\quad \boldsymbol{m}=(m_1,\ldots,m_d)\in\{0,1,\ldots,N\}^d.$$

Note that the $\Phi_{\boldsymbol{m}}^{(\alpha,\beta)}$ give rise to an orthonormal basis of $L^2(I^d,\boldsymbol{\omega}_{\alpha,\beta})$, where $\boldsymbol{\omega}_{\alpha,\beta}(\boldsymbol{x})=\prod\limits_{j=1}^{d}\omega_{\alpha,\beta}(x_j)$. For reasons of simplicity and readability, we restrict ourselves to the case $d=1$ in this work.

In the second part of this work, we are interested in the construction of a random pseudo-inverse based estimator for solving linear functional regression (LFR) problem. Recall that for a compact interval J, the LFR model is given as follows, see for example [11],

$$Y_i=\int_J X_i(s)\,\beta_0(s)\,dt+\varepsilon_i,\quad i=1,\ldots,n. \tag{3}$$

Here, Y_i is the scalar response, $X_i(\cdot)\in L^2(J)$ are random functional predictors, the ε_i are i.i.d centred white noise independent of the $X_i(\cdot)$ and $\beta_0(\cdot)\in L^2(J)$ is the unknown slope function to be recovered. We assume that $X_i(\cdot)$ lies in a finite dimensional subspace $\mathcal{H}_N$ of $L^2(J)$ and it is given by $X_i(s)=\sum\limits_{k=1}^{N}\xi_k Z_{i,k}\varphi_k(s)$, where $\{\varphi_k(\cdot),\, k=1,\ldots,N\}$ is an orthonormal family of $L^2(J)$, the unknown slope function $\beta_0(\cdot)$ is given (or efficiently approximated) by an expansion with respect to possibly another orthonormal set $\{\psi_j(\cdot),\, 1\leq j\leq M\}$ of $L^2(J)$. that is $\beta_0(s)=\sum_{j=1}^{M}d_j\psi_j(s), s\in J$. The $Z_{i,k}$ are i.i.d centred random variables with variance σ_Z^2 and $(\xi_k)_{1\leq k\leq N}$ is a finite deterministic sequence of $\mathbb{R}\setminus\{0\}$. Typically, we assume that $M\leq N$ in the sense that the second basis is better adapted for the approximation of $\beta_0(\cdot)$. By using the previous

expansions, it is easy to see that a random pseudo-inverse estimator for the approximate solution of (3) is given by

$$\widehat{\beta}_{n,M}(s) = \sum_{j=1}^{M} \widehat{d}_j \psi_j(s), \qquad \widehat{\boldsymbol{d}} = [\widehat{d}_1, \ldots, \widehat{d}_M]' = \mathcal{G}_M^{-1} \cdot \Big(\mathcal{F}_M' \cdot \frac{1}{\sqrt{n}} [Y_i]'_{1 \le i \le n} \Big). \tag{4}$$

Here, the reduced size, positive definite $N \times N$ random matrix G_M is given by

$$\mathcal{G}_M = \mathcal{F}_M' \mathcal{F}_M, \quad \mathcal{F}_M = F_N \cdot T_{N,M}, \qquad F_N = \frac{1}{\sqrt{n}} \Big[\xi_j Z_{i,j} \Big]_{1 \le i \le n, 1 \le j \le N}, \quad T_{N,M} = \Big[< \varphi_k(\cdot), \psi_j(\cdot) > \Big]_{\substack{1 \le k \le N \\ 1 \le j \le M}} \tag{5}$$

We assume that the transformation matrix $T_{N,M} \in \mathbb{R}^{N \times M}$ has a $2-$condition number $\kappa_2(T_{N,M})$ of reasonable magnitude. It is well known that since $n \ge N \ge M$, then $\kappa_2(\mathcal{F}_M) = \kappa_2(F_N \cdot T_{N,M}) \le \kappa_2(F_N)\kappa_2(T_{N,M})$, so that

$$\kappa_2(\mathcal{G}_M) \le \kappa_2(G_N) \big(\kappa_2(T_{N,M}) \big)^2. \tag{6}$$

Next, we let $\boldsymbol{\xi}(\cdot)$ denote the function defined on J by $\boldsymbol{\xi}(s) = \sum_{k=1}^{N} \xi_k \varphi_k(s)$ and let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Assume that $\max_{k \ge 1}(|Z_k|) \le M_Z$ almost surely. Then, under the previous notations and assumptions, for any $\eta > 0$, we have with probability at least $1 - 2\exp\left(\dfrac{-n\eta^2}{2M_{\boldsymbol{\xi},N}^2} \right)$,

$$\kappa_2(G_N) \le \frac{1.72 \max_{k \ge 1} \sigma_Z^2 \xi_k^2 + \frac{M_{\boldsymbol{\xi},N}}{n} \log(N) + \eta}{0.63 \min_{k \ge 1} \sigma_Z^2 \xi_k^2 - \frac{M_{\boldsymbol{\xi},N}}{n} \log(N) - \eta}. \tag{7}$$

Here, $M_{\boldsymbol{\xi},N} = M_Z^2 \max\limits_{1 \le j \le N} |\xi_j| \cdot \|\boldsymbol{\xi}\|_{l_1}$.

The following theorem provides us with the squared L^2-error of our estimator $\widehat{\beta}_{n,N}(\cdot)$ of $\beta_0(\cdot)$.

Theorem 2. *Assume that* $\max\limits_{k \ge 1}(|Z_k|) \le M_Z$ *almost surely and let* $\gamma_k^2 = \Big(\sum\limits_{j=1}^{M} < \varphi_k(\cdot), \psi_j(\cdot) > d_j \Big)^2$. *Then, under the previous notations and assumptions, for any* $\eta > 0$, *the following squared* L^2-*error of our estimator* $\widehat{\beta}_{n,M}$

$$\|\widehat{\beta}_{n,M}(\cdot) - \beta_0(\cdot)\|_{L_2}^2 = \|\widehat{\boldsymbol{d}}_{n,N} - \boldsymbol{d}\|_{\ell_2}^2 \le \kappa_2(\mathcal{G}_M) \frac{\frac{1}{n}\|\boldsymbol{\varepsilon}\|_{\ell_2}^2 \|\beta_0(\cdot)\|_{L_2}^2}{\sigma_Z^2 \max_{k \ge 1} \xi_k^2 \gamma_k^2 - \eta}, \tag{8}$$

holds with probability at least $1 - \exp\left(-\dfrac{2n\,\eta^2}{M_Z^4 \|\boldsymbol{\xi}(\cdot)\|_{\ell_2}^4 \|\beta_0(\cdot)\|_{L_2}^4} \right)$.

Remark 2. *When* $\beta_0(.)$ *and* $X_i(\cdot)$ *are expanded in the same basis* $\{\varphi_k(\cdot),\, k = 1, \ldots, N\}$, *then we have* $T_{N,N} = I$ *and so*

$$\mathcal{G}_N = \mathcal{F}_N' \mathcal{F}_N, \quad , \qquad F_N = \frac{1}{\sqrt{n}} \Big[\xi_j Z_{i,j} \Big]_{1 \le i \le n, 1 \le j \le N}$$

Note that the squared L^2-*error, given by* (8) *is trivially adapted to the estimator* $\widehat{\beta}_{n,N}(\cdot)$, *given by* (4). *To this end, it suffices to replace* γ_k^2 *by* d_k^2. *Then, under the previous notations and assumptions, one gets for any* $\eta > 0$,

$$\|\widehat{\beta}_{n,N}(\cdot) - \beta_0(\cdot)\|_{L_2}^2 = \|\widehat{\boldsymbol{d}}_{n,N} - \boldsymbol{d}\|_{\ell_2}^2 \le \kappa_2(\mathcal{G}_N) \frac{\frac{1}{n}\|\boldsymbol{\varepsilon}\|_{\ell_2}^2 \|\beta_0(\cdot)\|_{L_2}^2}{\sigma_Z^2 \max_{k \ge 1} \xi_k^2 d_k^2 - \eta}.$$

Numerical simulation In this paragraph, we give two numerical examples that illustrate the different results of this work.

Example 1: In this example, we illustrate the accuracy of our two estimators $\widehat{f}_{n,N}^{\alpha,\beta}$ and $\widehat{f}_n^{\lambda}$ in the case of a regression function belonging to the Sobolev space $H^s(I)$. We consider for f the Brownian motion function $f^s(x)$ given by:

$$f^s(x) = \sum_{k\geq 1} \frac{X_k}{k^s} \cos(k\pi x), \quad -1 \leq x \leq 1, \tag{9}$$

where $s > \frac{1}{2}$ is a positive real number and the X_k's are standard Gaussian random variables. It is well known that $f^s \in H^{s'}(I)$ almost surely for any $s' < s - \frac{1}{2}$. Here, $H^s(I)$ denotes the usual Sobolev space over I with Sobolev regularity exponent s. The random noise terms η_i are taken as $\eta_i = 0.1Z_i$ where Z_i follows the standard normal distribution with mean zero and variance 1. We consider two values for s, namely $s = 1, 2$. Then, we calculate $\widehat{f^s}_{n,N}^{\alpha,\alpha}$ with $\alpha = 0$, $N = 20$, $n = 100, 500, 1000$. For comparison purpose, we have computed the well known Kernel Ridge Regression estimator $\widehat{f}_n^{\lambda}$, associated with the previous Jacobi kernel. For this last estimator, we considered the values of $n = 50, 100, 150$. The regularization parameter λ is chosen by the GCV rule and it is equal to 0.01. We simulate 500 samples and we obtain the average (over the 500 samples) $L^2(I)$-regression error for each estimator. These results are reported in Table 1.

s	n	$\|f^s - \widehat{f}_{n,N}^{\alpha,\beta}\|^2$	n	$\|f^s - \widehat{f_n^{\lambda}}\|^2$
1	100	2.84e − 01	50	4.72e − 01
	500	2.34e − 01	100	3.13e − 01
	1000	2.27e − 01	150	2.88e − 01
2	100	1.12e − 01	50	1.64e − 01
	500	4.54e − 02	100	8.34e − 02
	1000	3.65e − 02	150	6.48e − 02

Table 1: Squared L^2-regression errors of example 1.

Example 2: In this example, we illustrate the results of Theorem 2 concerning the stability and the squared $L^2(I)-$errors of our estimator $\widehat{\beta}_{n,N}(\cdot)$ for the slope function, associated with the LFR problem. For this purpose, we use the following simulation test initially given in [6]. The interval $J = [0, 1]$ and the slope function is given by

$$\beta_0(s) = \sum_{j=1}^{50} 4\frac{(-1)^{j+1}}{j^2}\varphi_j(s), \quad \varphi_j(s) = \begin{cases} 1 & \text{if } j = 1 \\ \sqrt{2}\cos(\pi(j-1)s) & \text{if } j \geq 2. \end{cases}$$

The random predictor functional is given by $X(\cdot) = \sum_{k=1}^{50} \xi_k Z_k \varphi_k(\cdot)$, $\xi_k = \frac{(-1)^{k+1}}{k^{s/2}}$, $s \geq 0$, where the Z_k are independent samples following the uniform law $U(-\sqrt{3}, \sqrt{3})$. we illustrate the performance of the more general LFR estimator $\widehat{\beta}_{n,M}(\cdot)$, given by (4) and (5). For this purpose, we use the following expansion $\widehat{\beta}_{n,M}(\cdot)$ with respect to the first $M = 5$ orthonormal Legendre polynomials over $J = [0, 1]$, that is $\widehat{\beta}_{n,M}(s) = \sum_{j=1}^{5} \widehat{d}_j \sqrt{2j-1} P_{j-1}(2s-1)$, $s \in J$, where P_j is the usual Legendre polynomial defined on $[-1, 1]$ with the normalization $P_j(1) = (-1)^j$. Then, we have applied the estimator $\widehat{\beta}_{n,M}$ to the previous LFR problem with fixed $N = 50$ and different values of $100 \leq n \leq 500$. Table 2 lists the averages over 100 realizations of the squared L^2-errors associated with the estimator $\widehat{\beta}_{n,M}(\cdot)$. From these numerical simulations, one concludes that the general estimator $\widehat{\beta}_{n,M}(\cdot)$ is fast, accurate and it is particularly adapted for the LFR problem considered in this example.

s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|^2_{L_2}$	s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|^2_{L_2}$	s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|^2_{L_2}$
0.25	100	2.95e-2	1.1	100	3.45e-2	2	100	5.28e-1
–	200	1.25e-2	–	200	2.54e-2	–	200	3.94e-2
–	300	8.22e-3	–	300	1.25e-2	–	300	2.88e-2
–	500	1.54e-3	–	500	7.32e-3	–	500	1.38e-2

Table 2: Squared L^2-errors associated with the estimator $\widehat{\beta}_{n,M}$ with $M = 5$ and different values of s, n.

References

[1] A. Bonami and A. Karoui, Random Discretization of the Finite Fourier Transform and Related Kernel Random Matrices, *J. Fourier Anal. Appl.,* **26** 29 (2020), Doi: 10.1007/s00041-020-09736-8.

[2] Bousselmi B., Dupuy J. P. and Karoui A., Random orthogonal projections based schemes for solving nonparametric and linear functional regressions problems, submitted for publication, (2020).

[3] T. Cai, and P. Hall, Prediction in functional linear regression, *Ann. Statist.,* **34** (2006), 2159–2179.

[4] F. Comte and V. G. Catalot, Regression function estimation as a partly inverse problem, *Ann. Inst. Stat. Math.,* **72** (2020), 1023–1054.

[5] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.,* **39** (1) (2002), 1–49.

[6] P. Hall and J. L. Horowitz, Methodology and convergence rates for functional linear regression, *Ann. Statist.,* **35** (1), (2007), 70–91.

[7] I. Pinelis, An approach to inequalities for the distributions of infinite-dimensional martingales, In: Dudley R.M., Hahn M.G., Kuelbs J. (eds) Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference. Progress in Probability, **30** Birkhäuser, Boston, MA, (1992) 128–134.

[8] S. Smale and D. X. Zhou, Shannon sampling II: Connections to learning theory, *Appl. Comput. Harmon. Anal.* **19** (2005), 285–302.

[9] J. A. Tropp, An Introduction to Matrix Concentration Inequalities, Foundations and Trends in Machine Learning series, **8** No. 1–2, Now Publishers Inc., (2015).

[10] E. De Vito, A. Caponnetto and L. Rosasco, Model Selection for Regularized Least-Squares Algorithm in Learning Theory, *Found. Comput. Math.,* **5** (2005), 59–85.

[11] M. Yuang and T. T. Cai, A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression, *Ann. Stat.,* **38** (6) (2010), 3412–3444.

Pénalisation l_1 pour un mélange de lois de von Mises-Fisher

Florian Barbaro [1] & Fabrice Rossi [2]

[1] *Université Paris 1 Panthéon-Sorbonne - Laboratoire SAMM EA 4543, florian.barbaro@etu.univ-paris1.fr*
[2] *Université Paris Dauphine-PSL - CEREMADE UMR 7534, rossi@ceremade.dauphine.fr*

Résumé. Les mélanges de lois de von Mises-Fisher permettent de construire des classifications (non supervisées) de données sur la sphère unité. Ces mélanges sont bien adaptés aux données directionnelles de grande dimension comme les textes. Pour améliorer la qualité des classes et leur interprétabilité, nous proposons dans cet article de pénaliser la vraisemblance par un terme l_1, ce qui conduit à des centroïdes parcimonieux. Nous dérivons un algorithme EM pour ce modèle et nous illustrons l'intérêt de notre approche sur un jeu de données réelles.

Mots-clés. Mélanges de lois de von Mises-Fisher, pénalisation l_1, données de grande dimension.

Abstract. Mixtures of von Mises-Fisher distributions can be used to cluster data on the unit hypersphere. This is particularly adapted for high dimensional directional data such as texts. We propose in this article to estimate a von Mises mixture using a l_1 penalised likelihood. This leads to sparse prototypes that improve both clustering quality and interpretability. We introduce an EM algorithm for this estimation and show the advantages of the approach on real data benchmark.

Keywords. Mixtures of von Mises-Fisher distributions, l_1 penalty, high-dimensional data.

1 Introduction

Beaucoup de modèles de mélanges classiques sont peu adaptés aux données de grande dimension, par exemple issues de la représentation vectorielle de textes. Quand les données sont directionnelles [Mardia and Jupp, 2009], c'est-à-dire quand c'est plutôt leur corrélation que leur distance euclidienne qui importe, les modèles de type Gaussien sont encore moins adaptés. Pour de telles données, il est naturel d'opérer à une normalisation qui les place sur la sphère unité. On montre alors que les mélanges de lois de von Mises-Fisher (vMF) sur cette sphère sont bien adaptées pour la classification (non supervisée), cf [Banerjee et al., 2005, Gopal and Yang, 2014].

Dans cet article, en s'inspirant de [Pan and Shen, 2007], nous proposons une pénalisation l_1 pour un mélange de distributions vMF pour augmenter la parcimonie des moyennes

directionnelles et ainsi améliorer la compréhension des résultats de classification des données de grande dimension. Notre solution s'appuie sur une modification de l'algorithme espérance-maximisation proposé par [Banerjee et al., 2005].

Notations Les matrices sont indiquées en gras et majuscules, les vecteurs en minuscules et en gras. La norme l_1 est notée par $\|.\|_1$ et la norme l_2 par $\|.\|_2$. La sphère unité de dimension $(d-1)$ intégrée dans $\mathbb{R}^d$ est noté $\mathbb{S}^{d-1}$. Les données sont représentées par une matrice $\boldsymbol{X} = (x_{ij})$ de dimension $n \times d$ avec $x_{ij} \in \mathbb{R}$ et la i^{eme} ligne de cette matrice est représentée par un vecteur $\boldsymbol{x_i} = (x_{i1}, ..., x_{id})^T$, où T dénote la transposée. La partition de l'ensemble des lignes I en K classes peut être représentées par une matrice de classification $\boldsymbol{Z}$ d'éléments z_{ih} dans $\{0, 1\}$ satisfaisant $\sum_{h=1}^{K} z_{ih} = 1$. On note $\mathbb{I}$ la fonction caractéristique.

2 Mélange de lois de von Mises-Fisher

On rappelle tout d'abord le modèle de mélange proposé dans [Banerjee et al., 2005]. La densité de la loi de von Mises en un point $\boldsymbol{x_i} \in \mathbb{S}^{d-1}$ est donnée par

$$f(\boldsymbol{x_i}|\boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^T \boldsymbol{x_i}\right). \tag{1}$$

où $\boldsymbol{\mu}$ est la moyenne directionnelle et κ le paramètre de concentration de la loi, tels que $\|\boldsymbol{\mu}\|_2 = 1$ et $\kappa \geq 0$. Le terme de normalisation $C_d(\kappa)$ est donné par $C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ où I_r est la fonction de Bessel modifiée du premier type d'ordre r.

On considère un mélange de K lois de van Mises, chacune avec ses propres paramètres, avec la densité [Banerjee et al., 2005]

$$f(\boldsymbol{x_i}|\boldsymbol{\Theta}) = \sum_{h=1}^{K} \alpha_h f(\boldsymbol{x_i}|\boldsymbol{\mu_h}, \kappa_h). \tag{2}$$

où $\Theta = \{\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \alpha_1, \ldots, \alpha_K, \kappa_1, \ldots, \kappa_K\}$. Les observations sont supposées indépendantes. En introduisant les variables latentes $\boldsymbol{Z}$ indiquant (sous forme des indicatrices z_{ih}) la composante du mélange responsable de chaque observation, on obtient la log-vraisemblance suivante pour les données complétées :

$$l(\Theta|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{h=1}^{K} z_{.h}[\log \alpha_h + \log c_d(\kappa_h)] + \sum_{i,h} z_{ih} \kappa_h \boldsymbol{\mu_h}^T \boldsymbol{x_i}, \tag{3}$$

où $z_{.h}$ représente la cardinalité de la classe h. En s'appuyant sur cette vraisemblance complétée, l'utilisation d'un algorithme EM pour l'estimation des paramètres ne pose pas de problème spécifique, excepté l'estimation des κ_h, cf [Banerjee et al., 2005] pour des détails.

3 Modèle proposé

3.1 Vraisemblance pénalisée

Nous proposons de pénaliser la vraisemblance par la norme l_1 permettant ainsi d'augmenter la parcimonie de la représentation des moyennes directionnelles. Plus précisément, nous cherchons à estimer Θ en maximisant la log-vraisemblance pénalisé :

$$l_p(\Theta|\boldsymbol{X}) = l(\Theta|\boldsymbol{X}) - \beta \sum_{h=1}^{K} \|\boldsymbol{\mu_h}\|_1 , \tag{4}$$

où β règle le compromis entre la vraisemblance et la parcimonie. Comme le montre [Pan and Shen, 2007], cette pénalisation n'a pas d'effet sur l'étape E de l'algorithme EM pour un modèle de mélange.

3.2 Phase M de l'algorithme EM

En revanche, la phase est modifiée. Notons $\tau'_{i,h} = \mathbb{P}(z_{ih} = 1|\boldsymbol{x_i}, \Theta')$, où Θ' désigne l'estimation actuelle des paramètres. L'espérance par rapport à $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}, \Theta')$ de la log-vraisemblance pénalisée s'écrit alors

$$Q_P(\Theta|\Theta') = \sum_{h=1}^{K} \tau'_{.h}[\log \alpha_h + \log c_d(\kappa_h)] + \sum_{i,h} \tau'_{ih} \kappa_h \boldsymbol{\mu_h}^T \boldsymbol{x_i} - \beta \sum_{h=1}^{K} \|\boldsymbol{\mu_h}\|_1 , \tag{5}$$

où $\tau'_{.h} = \sum_i \tau'_{ih}$. On introduit le Lagrangien

$$\mathcal{L}(\Theta, \boldsymbol{\lambda}|\Theta') = Q_P(\Theta|\Theta') + \sum_h \lambda_h (1 - \boldsymbol{\mu_h}^T \boldsymbol{\mu_h}). \tag{6}$$

Par rapport aux dérivations de [Banerjee et al., 2005], la différence principale vient du calcul du sous-gradient de $\mathcal{L}$ par rapport à $\mu_{h,j}$. On a en effet

$$\partial_{\mu_{hj}} \mathcal{L}(\Theta, \boldsymbol{\lambda}|\Theta') = \kappa_h \sum_i \tau'_{ih} x_{ij} - 2\lambda_h \mu_{hj} - \beta \partial_{\mu_{hj}} |\mu_{hj}|. \tag{7}$$

Dans la dérivation qui suit, nous nous restreignons au cas où les μ_{hj} sont positifs ou nuls, pour une application à des données positives de type textes. Cette dérivation s'étend sans difficulté au cas général.

La condition d'optimalité du premier ordre est $0 \in \partial_{\mu_{hj}} \mathcal{L}(\Theta, \boldsymbol{\lambda}|\Theta')$. En notant $r'_{hj} = \sum_i \tau'_{ih} x_{ij}$, on a :

$$\partial_{\mu_{hj}} \mathcal{L}(\Theta, \boldsymbol{\lambda}|\Theta') = \begin{cases} \kappa_h r'_{hj} - 2\lambda_h \mu_{hj} - \epsilon\beta, \epsilon \in [-1; 1] & \text{si } \mu_{hj} = 0 \\ \kappa_h r'_{hj} - 2\lambda_h \mu_{hj} - \beta & \text{si } \mu_{hj} > 0 \end{cases} \tag{8}$$

On en déduit que $\mu_{hj} = \max\left(\frac{\kappa_h r'_{hj} - \beta}{2\lambda_h}, 0\right)$. En réinjectant cette formule dans la contrainte $\|\boldsymbol{\mu}_h\|_2 = 1$, on trouve

$$\lambda_h = \frac{1}{2}\sqrt{\sum_j (\max(\kappa_h r'_{hj} - \beta, 0))^2}., \tag{9}$$

ce qui permet de conclure que

$$\mu_{hj} = \max\left(\frac{\kappa_h r'_{hj} - \beta}{\sqrt{\sum_l (\max(\kappa_h r'_{hl} - \beta, 0))^2}}, 0\right). \tag{10}$$

Notons que l'ajout de la pénalisation introduit un couplage entre κ_h et $\boldsymbol{\mu}_h$ qui n'existe pas en son absence (on voit que si on fixe $\beta = 0$, κ_h n'intervient plus dans la définition de $\boldsymbol{\mu}_h$). On doit donc résoudre

$$\frac{c'_d(\kappa_h)}{c_d(\kappa_h)} = -\frac{\boldsymbol{\mu}_h \sum_i \tau'_{ih} \boldsymbol{x}_i}{\sum_i \tau'_{ih}}. \tag{11}$$

Nous reprenons l'approximation proposée dans [Banerjee et al., 2005]. Si on pose $\bar{r}'_h = \frac{\boldsymbol{\mu}_h \sum_i \tau'_{ih} \boldsymbol{x}_i}{\sum_i \tau'_{ih}}$, on estime κ_h par

$$\kappa_h = \frac{\bar{r}'_h d - (\bar{r}'_h)^3}{1 - (\bar{r}'_h)^2}. \tag{12}$$

On propose d'estimer $\boldsymbol{\mu}_h$ à partir de κ'_h, puis de mettre à jour κ_h.

3.3 Sélection de modèle

Nous proposons de sélectionner le modèle retenu pour un jeu de données en utilisant le critère BIC. Seuls les paramètres non nuls pour μ_{hj} sont considérés comme des paramètres effectifs. On a donc

$$BIC = -2 \times l(\hat{\Theta}|\boldsymbol{X}) + C \times \log(n), \tag{13}$$

avec pour C le nombre de paramètres la valeur $C = (K - 1 + K) + \sum_h \sum_j \mathbb{I}_{\mu_{hj} \neq 0}$.

4 Résultats expérimentaux

Pour obtenir les résultats expérimentaux, les équations 9, 10 et 13 sont implémentées à l'aide du *package* R *movMF*[1]. L'algorithme est ainsi testé sur un jeu de données textuelles en comparaison avec le modèle initial. Pour comparer les modèles nous avons choisi d'utiliser le Adjusted Rand Index.

1. https ://cran.r-project.org/web/packages/movMF/index.html

Nous avons sélectionné un jeu de données populaires pour tester notre algorithme à savoir CSTR [Li, 2005] [2] avec les caractéristiques suivantes $(n, d, g) = (475, 1000, 4)$. Il est composé de résumés de rapports techniques (TR) publiés au Département d'informatique de l'Université de Rochester entre 1991 et 2002. De plus, il a été divisé en quatre catégories qui sont *Natural Language Processing(NLP)*, *Robotics/Vision*, *Systems*, et *Theory*.

Pour commencer l'analyse, il est intéressant de s'attarder sur les modèles sélectionnés par le BIC. Pour le modèle movMF original, le BIC a été calculé sur le modèle qui maximise la vraisemblance pour chaque classe. Pour celui pénalisé, les méta-paramètres β et K sont estimés avec le BIC. Les résultats sont visibles sur les figures 1 et 2 où l'on remarque que le BIC a sélectionné dans les deux cas, les modèles avec quatre classes. De plus, sur la figure 2, on distingue que les β sont très différents selon les modèles et que la parcimonie évolue en conséquence où elle atteint un maximum pour le modèle sélectionné.

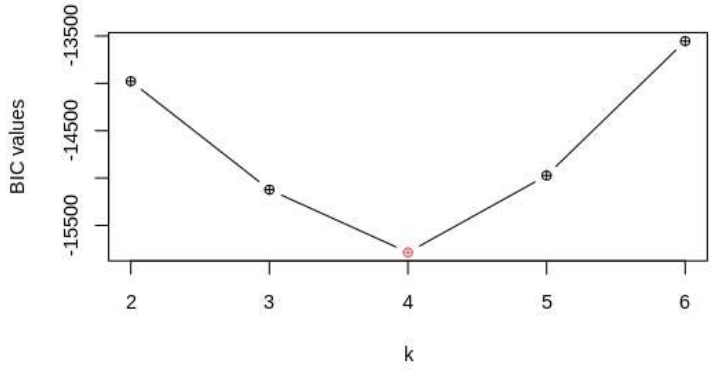


FIGURE 1 – Valeur du BIC pour movMF selon k.

La table 1 montre quant à elle les résultats obtenus avec les modèles sélectionnés par le BIC. Le modèle pénalisé avec un $\beta = 142$ obtient un ARI supérieur au movMF et permet une grande parcimonie de la moyenne directionnelle.

TABLE 1 – Résultats.

Algo	ARI	Parcimonie
movMF	63%	0%
movMF pénalisé	**72%**	**67%**

2. Disponible ici : https ://github.com/dbmovMFs/DirecCoclus/tree/master/Data

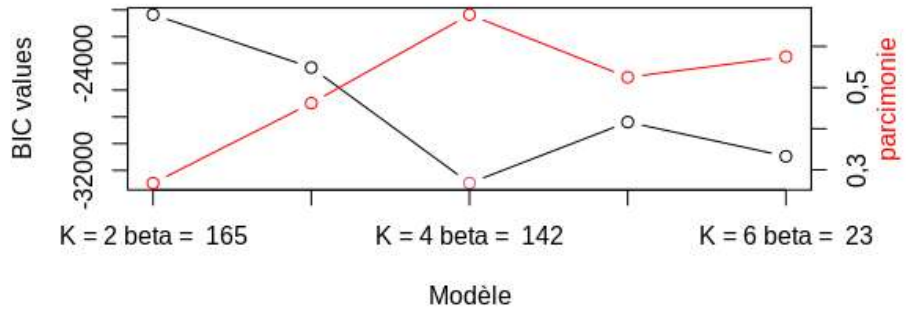


FIGURE 2 – Valeurs du BIC et de la parcimonie pour movMF pénalisé selon k et β.

Références

[Banerjee et al., 2005] Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46) :1345–1382.

[Gopal and Yang, 2014] Gopal, S. and Yang, Y. (2014). Von mises-fisher clustering models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 154–162, Bejing, China. PMLR.

[Li, 2005] Li, T. (2005). A general model for clustering binary data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 188–197, New York, NY, USA. Association for Computing Machinery.

[Mardia and Jupp, 2009] Mardia, K. and Jupp, P. (2009). *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley.

[Pan and Shen, 2007] Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(41) :1145–1164.

Statistical properties of functional principal components analysis based on discretized observations

Ryad Belhakem [1] & Franck Picard [2] & Vincent Rivoirard [3] & Angelina Roche [4]

[1],[3],[4] *CEREMADE Université Paris-Dauphine Place du Maréchal de Lattre de Tassigny 75775 PARIS CEDEX 16*

[2] *Laboratoire de Biologie et Modélisation de la Cellule ENS de Lyon 46, allée d'Italie 69364 LYON CEDEX 07*

[1] *belhakem@ceremade.dauphine.fr*

[2] *franck.picard@ens-lyon.fr*

[3] *rivoirard@ceremade.dauphine.fr*

[4] *roche@ceremade.dauphine.fr*

Abstract. Functional principal components analysis relies on a preliminary smoothing step using a projection on smooth functional basis. The underlying assumption at the core of this common practice consists in neglecting the noise at fine-scales. However, when eigen-functions are non-smooth or smoother at a fine scales, the projection step might worsen the results. The aim of the present contribution is to investigate theoretical properties of estimators constructed by using this preliminary smoothing step, for general basis and for the ones that are best fit for irregular signals such as histograms and Haar wavelets.

Keywords. Functional principal components analysis, high-dimensional statistics, Wavelets.

Abstract. L'analyse en composantes principales fonctionnelles repose souvent sur une étape préalable de lissage. Celle-ci se fait par le biais d'une projection sur une base de fonctions régulières. Cette pratique implique une certaine régularité du signal. De fait, les variations rapides ou sur de petites échelles sont considérées comme du bruit. Lorsque les fonctions propres ne sont pas lisses ou lisses à une échelle plus fine, l'étape de projection peut détériorer les résultats. Le but de la présente contribution est d'étudier les propriétés théoriques des estimateurs construits au moyen de cette étape préalable de lissage, pour une base générale et pour celles qui conviennent le mieux pour des signaux irréguliers tels que les histogrammes et les ondelettes d'Haar.

Keywords. Analyse en composantes principales fonctionnelle, statistique en grande dimension, Ondelettes.

1 Functional principal components analysis

We observe i.i.d replicates of Y, a noisy version a random function $Z \in L^2([0,1])$, such that $\mathbf{E}[Z] = 0$ and $\mathbf{E}[\|Z\|_{L^2([0,1])}] < \infty$. This function is observed on a fixed grid $\{t_h = h/p; \quad h = 0, \dots, p-1\}$ such that $\{Y_i(t_h) = Z_i(t_h) + \epsilon_{i,h}, i = 1, \dots, n, h = 0, \dots, p-1\}$. We suppose that $\epsilon_{i,h} \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$ and that errors are independent from Z. In the following, we denote by K_Z the covariance function of Z defined such that:

$$\forall (s,t) \in [0,1]^2, \quad K_Z(s,t) \;=\; \mathbf{E}(Z(s)Z(t)),$$

Thanks to the Karhunen-Loève decomposition (see Bosq), such that:

$$\forall t \in [0,1] \quad Z(t) = \sum_{\ell \in N} \zeta_\ell (\mu_\ell^*)^{1/2} \eta_\ell^*(t),$$

with $(\zeta_\ell)_{\ell \in N}$ a sequence of non-correlated centered random variables of variance 1, and $\mu_1^* > \mu_2^* > \dots$ stand for the ordered eigenvalues, supposed to be distinct, with corresponding eigenfunctions $(\eta_\ell^*)_{\ell \in N}$. The purpose of fPCA is to estimate this functional eigen decomposition.

This Gaussian process model is studied in Ramsay and Silverman [1] and Li and Hsing [2]. The standard approach for fPCA consists in smoothing the observation using splines and apply FPCA to the smoothed observations. Yao et al [3] developed an approach that consists in neglecting the diagonal term of the covariance, using splines to estimate the diagonal and reconstruct the kernel, and applying FPCA on the subsequent results. Descary and Panaretos [4] generalized this approach to a non i.i.d noise using stronger regularity conditions. If assuming that Y_i was observed over the whole interval $[0,1]$, theoretical results exist (such as Mas and Ruymgaart [5]), with estimators attaining minmax rates. However, in practice a curve is observed on a finite grid. This results in a discretization of the functional data which impacts the estimator's risk, with a combined asymptotic behavior in (n,p) the number of observations and the size of the observational grid.

1.1 Preliminary smoothing step

We denote by $(\phi_\lambda)_{\lambda \in \Lambda}$ an orthonormal basis of $L^2[0,1]$ that is used to smooth the data, such that, by setting $\beta_\lambda := \langle Y, \phi_\lambda \rangle$,

$$\forall t \in [0,1] \quad Y(t) = \sum_{\lambda \in \Lambda} \langle Y, \phi_\lambda \rangle \phi_\lambda(t) = \sum_{\lambda \in \Lambda} \beta_\lambda \phi_\lambda(t), \tag{1}$$

and we define for any $(\lambda, \lambda') \in \Lambda^2$:

$$\gamma_{\lambda,\lambda'} := \mathbf{E}[\langle Y, \phi_\lambda \rangle \langle Y, \phi_{\lambda'} \rangle] = \mathbf{E}[\beta_\lambda \beta_{\lambda'}] = \iint_{[0,1]^2} K_Y(t,s) \phi_\lambda(t) \phi_{\lambda'}(s) dt ds.$$

To define the estimators of the functional principal components (fPCs), we introduce the empirical quantities of our model. To investigate the effect of the discretization on the perfomance of our method, we will pay a particular attention to notations : for a given quantity, say γ, $\widetilde{\gamma}$ will refer to its discretized theoretical version and, $\widehat{\widetilde{\gamma}}$ will refer to its empirical discretized version. For instance, when p and n are large, coefficients $\gamma_{\lambda,\lambda'}$ are close to their discretized empirical versions defined by:

$$\widehat{\widetilde{\gamma}}_{\lambda,\lambda'} := \frac{1}{np^2}\sum_{h,h'=0}^{p-1}\sum_{i=1}^{n} Y_i(t_h)Y_i(t_{h'})\phi_\lambda(t_h)\phi_{\lambda'}(t_{h'}).$$

Let $\Lambda_p \subset \Lambda$ such that $|\Lambda_p| < \infty$. We denote by $\widehat{\widetilde{G}}_\phi$ the matrix of size $|\Lambda_p|^2$ with components:

$$\widehat{\widetilde{G}}_{\phi,\lambda,\lambda'} := \widehat{\widetilde{\gamma}}_{\lambda,\lambda'}.$$

Since we only have access to a finite number $|\Lambda_p|^2$ of coefficients, the optimization problem resumes is finite, which is exactly equivalent to a PCA problem, where the matrix to diagonalize is $\widehat{\widetilde{G}}_\phi$. We estimate then $(a_1^*, \ldots, a_d^*, \mu_1^*, \ldots, \mu_d^*)$ as a solution of the following optimization problem:

$$(\widehat{\widetilde{a}}_1, ..., \widehat{\widetilde{a}}_d, \widehat{\widetilde{\mu}}_1, \ldots, \widehat{\widetilde{\mu}}_d) \quad = \quad \underset{\substack{\langle a_\ell, a_{\ell'}\rangle = \delta_{\ell,\ell'} \\ \mu_\ell \in \mathbf{R}_+}}{\arg\min} \; \|\widehat{\widetilde{G}}_\phi - \sum_{\ell=1}^{d} \mu_\ell a_\ell^T a_\ell\|_F^2, \tag{2}$$

We finally $(\eta_1^*, ..., \eta_d^*)$ are estimated using:

$$\widehat{\widetilde{\eta}}_\ell = \sum_{\lambda\in\Lambda_p} \widehat{\widetilde{a}}_{\lambda,\ell}\phi_\lambda, \quad \ell = 1, \ldots, d. \tag{3}$$

2 Theoretical results :

Since our estimators are based on a Riemann approximation of the integral and not a Monte Carlo, we need to enforce stronger conditions of regularity on the observation and the basis used.

Assumption H1: $\exists \alpha \in]0,1[$, such that

$$\exists C_Z > 0, \forall (s,t) \in [0,1]^2 \quad \mathbf{E}[(Z(t) - Z(s))^2] \leq C_Z|t-s|^{2\alpha}. \tag{4}$$

To control the bias induced by the discretization of the basis we make a similar hypothesis, we suppose that functions $(\phi_\lambda)_{\lambda\in\Lambda}$ are Lipchitz continuous almost everywhere.

Assumption H2: $\exists (L_\lambda)_\lambda$ such that, almost everywhere,

$$\forall (s,t) \in [0,1]^2, \quad |\phi_\lambda(t) - \phi_\lambda(s)| \leq L_\lambda|t-s|. \tag{5}$$

Note that H2 covers all common bases such as wavelets (except the Haar basis), polynomial and the Fourier basis.

Lipschitz basis We assume that H1 and H2 are satisfied. Let $\ell \in \{1, \dots, d\}$ be fixed. Then, with $c_\ell = 8/\min(\mu_\ell - \mu_{\ell+1}, \mu_{\ell-1} - \mu_\ell)^2$, let $v = \max_{t\in[0,1]} \sqrt{\mathbf{E}[Z(t)^2]}$ we have with probability larger than $1 - 2p^2 \max\left\{\exp(-4\log(n)); \exp\left(-\sqrt{2(v^2+\sigma^2)n}\right)\right\}$

$$\begin{aligned}\frac{1}{c_\ell}\|\widehat{\widetilde{\eta}}_\ell - \eta_\ell^*\|_{L^2}^2 \leq{}& \frac{384\|\widetilde{\phi}\widetilde{\phi}^T\|_2^2(v^2+\sigma^2)\log(n)}{n} + \frac{3\sigma^4\|\widetilde{\phi}\widetilde{\phi}^T\|_2^2}{p} \\ &+ \frac{48L_K^2(\sum_{\lambda\in\Lambda_p}\|\phi_\lambda\|_{L^1}^2)^2}{p^{2\alpha}} + \sum_{(\lambda',\lambda)\in\Lambda_p^2}\frac{6L_{\lambda'}^2\|K_Z\|_\infty^2\|\phi_\lambda\|_{L^1}^2}{p^2} \\ &+ \frac{3(\sum_{\lambda\in\Lambda_p}L_\lambda^2)^2\|K_Z\|_\infty^2}{4p^4} + 3\sum_{(\lambda,\lambda')\notin\Lambda_p^2}\gamma_{\lambda,\lambda'}^2\end{aligned}$$

and

$$\begin{aligned}\sup_{\ell\geq 1}|\widehat{\widetilde{\mu}}_\ell - \mu_\ell^*|^2 \leq{}& \frac{384\|\widetilde{\phi}\widetilde{\phi}^T\|_2^2(v^2+\sigma^2)\log(n)}{n} + \frac{3\sigma^4\|\widetilde{\phi}\widetilde{\phi}^T\|_2^2}{p} \\ &+ \frac{48L_K^2(\sum_{\lambda\in\Lambda_p}\|\phi_\lambda\|_{L^1}^2)^2}{p^{2\alpha}} + \sum_{(\lambda',\lambda)\in\Lambda_p^2}\frac{6L_{\lambda'}^2\|K_Z\|_\infty^2\|\phi_\lambda\|_{L^1}^2}{p^2} \\ &+ \frac{3(\sum_{\lambda\in\Lambda_p}L_\lambda^2)^2\|K_Z\|_\infty^2}{4p^4} + 3\sum_{(\lambda,\lambda')\notin\Lambda_p^2}\gamma_{\lambda,\lambda'}^2.\end{aligned}$$

Histogram basis We assume H1 is satisfied. Let $v = \max_{t\in[0,1]} \sqrt{\mathbf{E}[Z(t)^2]}$ we have with probability larger than
$1 - 2p^2 \max\left\{\exp(-4\log(n)); \exp\left(-\sqrt{2(v^2+\sigma^2)n}\right)\right\}$:

$$\frac{1}{c_\ell}\|\widehat{\widetilde{\eta}}_\ell - \eta_\ell^*\|_{L^2}^2 \leq \frac{256(v^2+\sigma^2)\log(n)}{n} + \frac{8C_Y v^2}{p^{2\alpha}(2\alpha+1)}$$

$$\sup_{\ell\in\mathbf{N}^*}|\widehat{\widetilde{\mu}}_\ell - \mu_\ell^*|^2 \leq \frac{384(v^2+\sigma^2)\log(n)}{n} + \frac{12C_Y v^2}{p^{2\alpha}(2\alpha+1)} + \frac{3\sigma^4}{p}$$

Haar basis We assume H1 is satisfied and $p = 2^{k+1}$, fix $d \in \mathbf{N}$ the number of strongest eigenfunctions we want to estimate then $\forall \ell \in \{1, .., d\}$, fix $j_1 \in \mathbf{N}^*$ such that $j_1 = k$. Let $v = \max_{t\in[0,1]} \sqrt{\mathbf{E}[Z(t)^2]}$ we have with probability larger than
$1 - 2p^2 \max\left\{\exp(-4\log(n)); \exp\left(-\sqrt{2(v^2+\sigma^2)n}\right)\right\}$:

$$\frac{1}{c_\ell}\|\widehat{\widetilde{\eta}}_\ell - \eta_\ell^*\|_{L^2}^2 \leq \frac{256(v^2+\sigma^2)\log(n)}{n} + \frac{2C2^{2\alpha} + 8L_K^2(\log(2p)/\log(2))^2}{p^{2\alpha}}$$

$$\sup_{\ell \in \mathbf{N}} |\widehat{\widetilde{\mu}}_\ell - \mu_\ell^*|^2 \leq \frac{384(v^2+\sigma^2)\log(n)}{n} + \frac{3C2^{2\alpha} + 12L_K^2(\log(2p)/\log(2))^2}{p^{2\alpha}} + \frac{3\sigma^4}{p}$$

3 Discussion

Our results show that when the density of the grid increases (p tends to infinity), we achieve a parametric $O(n^{-1})$ rate of convergence. Functional PCA can be viewed as non-parametric learning problem based on the kernel $K(s,t) = \mathbf{E}[Y(t)Y(s)]$. Yao and al [3] showed that the error of estimation on the eigen-functions behaves as one dimensional non-parametric problem with rates of order $O(n^{\frac{-2\alpha}{2\alpha+1}})$. Faster rates are achieved in Descary and Panareto 2016 [6], under stronger conditions (assuming K to be analytic) they obtain a rate of order $O(n^{-1} + p^{-2})$, thus achieving parametric speeds when $p >> n$. This motivated our investigations and the goal was to have comparable results but under weaker conditions. However if analyticity is not assumed Descary and Panareto 2016 [6] showed that noise and signal are not distinguishable. To solve the problem we show that under i.i.d noise the effect of the noise is marginal, which leads to a rate of order $O(n^{-1} + p^{-2\alpha})$ on the reconstruction of eigenfunctions.

Bibliographie

[1] Ramsay, Jim and Silverman, B. W, 2010 *Springer New York* ,New York, NY.

[2] Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics* 3321–3351.

[3] Yao, F., Muller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100 577–590.

[4] Marie-Hélène Descary and Victor M. Panaretos, 2016. Functional Data Analysis by Matrix Completion, *The Annals of Statistics*

[5] André Mas, Frits Ruymgaart. High Dimensional Principal Projections. Complex Analysis and Operator Theory, *Springer Verlag*, 2015, 9 (1), pp.35 - 63.

[6] Marie-Hélène Descary and Victor M. Panareto, Functional Data Analysis by Matrix Completion, *The Annals of Statistics*, 2016.

MDA pour les forêts aléatoires : inconsistance, et une solution pratique via le Sobol-MDA

Clément Bénard [1,2] & Sébastien Da Veiga [2] & Erwan Scornet [3]

[1] *Sorbonne Université, CNRS, LPSM, 4 place Jussieu, 75005 Paris, France*
clement.benard@safrangroup.com
[2] *Safran Tech, Modeling & Simulation, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France*
[3] *CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, France*

Résumé. L'importance de variables est la principale approche pour décrypter le mécanisme "boîte noire" des forêts aléatoires. Bien que le *Mean Decrease Accuracy* (MDA) soit largement reconnu comme la mesure d'importance la plus efficace pour les forêts aléatoires, ses propriétés théoriques ont été peu explorées jusqu'à présent. En fait, la définition exacte du MDA varie d'une implémentation à une autre. C'est pourquoi nous formalisons mathématiquement les différentes implémentations du MDA, et établissons leur limite théorique lorsque la taille de l'échantillon augmente. Ces limites se décomposent en trois composantes : les deux premières sont liées aux indices de Sobol, qui sont des mesures bien définies de la contribution d'une variable d'entrée sur la variance de la sortie, et très utilisés en analyse de sensibilité, par opposition au troisième terme dont la valeur augmente avec la dépendance des entrées. Ainsi, nous démontrons théoriquement que le MDA n'estime pas la quantité théorique appropriée lorsque les entrées sont dépendantes, ce qui a déjà été observé empiriquement. Pour résoudre ce problème, nous introduisons une nouvelle mesure d'importance de variables pour les forêts aléatoires, le Sobol-MDA, qui corrige les défauts du MDA d'origine. Nous démontrons la consistance du Sobol-MDA, et illustrons ses bonnes performances empiriques à travers des expériences sur des données réelles et simulées. Une implémentation open source en `R` et `C++` est disponible en ligne.

Mots-clés. importance de variables, forêts aléatoires, MDA, analyse de sensibilité

Abstract. Variable importance measures are the main tools to analyze the black-box mechanism of random forests. Although the Mean Decrease Accuracy (MDA) is widely accepted as the most efficient variable importance measure for random forests, little is known about its theoretical properties. In fact, the exact MDA definition varies across the main random forest software. In this article, our objective is to rigorously analyze the behavior of the main MDA implementations. Consequently, we mathematically formalize the various implemented MDA algorithms, and then establish their limits when the sample size increases. In particular, we break down these limits in three components : the first two are related to Sobol indices, which are well-defined measures of a variable contribution to the output variance, widely used in the sensitivity analysis field, as opposed to the third term, whose value increases with dependence within input variables. Thus, we theoretically

demonstrate that the MDA does not target the right quantity when inputs are dependent, a fact that has already been noticed experimentally. To address this issue, we define a new importance measure for random forests, the Sobol-MDA, which fixes the flaws of the original MDA. We prove the consistency of the Sobol-MDA and show its good empirical performance through experiments on both simulated and real data. An open source implementation in `R` and `C++` is available online.

Keywords. variable importance, random forests, MDA, sensitivity analysis

1 Introduction

Forêts aléatoires. Les forêts aléatoires (Breiman, 2001) agrègent un grand nombre d'arbres pour effectuer des tâches de régression et de classification, et atteignent une grande précision pour un large éventail de problèmes. Cependant, les forêts souffrent d'un inconvénient majeur : un grand nombre d'opérations est calculé pour effectuer une prédiction, généralement des dizaines de milliers, ce qui rend impossible l'interprétation du mécanisme de prédiction. Cet aspect boîte noire est une forte limitation pratique, en particulier pour les applications impliquant des décisions critiques : le domaine médical ou l'optimisation des processus de fabrication en sont des exemples typiques. L'approche la plus répandue pour interpréter les forêts aléatoires est l'analyse d'importance de variables : les variables d'entrée sont classées par ordre décroissant de leur importance dans le processus de prédiction. Ainsi, des mesures d'importance spécifiques ont été développées pour les forêts aléatoires, et le MDA (Breiman, 2001) est le plus utilisé. Le MDA mesure l'augmentation de l'erreur lorsque les valeurs d'une variable d'entrée sont permutées, brisant ainsi sa relation avec la sortie : une valeur élevée de la métrique signifie que la variable est utilisée dans de nombreuses opérations du mécanisme de prédiction de la forêt. Malheureusement, il n'y a pas d'interprétation précise et rigoureuse puisque la définition du MDA est purement empirique. Notre objectif est d'utiliser l'analyse de sensibilité pour déterminer les propriétés théoriques du MDA, et d'introduire le Sobol-MDA qui corrige les défauts du MDA d'origine.

Analyse de sensibilité. Le but de l'analyse de sensibilité est de répartir les variations de la sortie d'un système entre les différentes entrées. En particulier, l'analyse de sensibilité globale (GSA) introduit des mesures bien définies de la contribution des entrées sur la variance de la sortie : les indices de Sobol, notamment utilisés pour analyser le comportement de codes numériques. Cependant, la littérature sur l'importance de variables en apprentissage statistique mentionne rarement l'analyse de sensibilité. Dernièrement, Gregorutti (2015) a établi un lien entre GSA et MDA : dans le cas d'entrées indépendantes, la contrepartie théorique du MDA est l'indice de Sobol total non normalisé. Nous étendons le lien entre les forêts aléatoires et l'analyse de sensibilité en démontrant la convergence

du MDA de Breiman dans le cas général d'entrées dépendantes et en décomposant la limite obtenue avec les indices de Sobol.

2 Définition du MDA

Une étude fine des principales implémentations des forêts aléatoires révèlent que plusieurs versions distinctes du MDA coexistent, sans formulation mathématique précise : le "Train-Test MDA" (TT-MDA), le "Breiman-Cutler MDA" (BC-MDA), et le "Ishwaran-Kogalur MDA" (IK-MDA). Pour formaliser ces définitions du MDA, nous commençons par introduire un cadre classique de régression, synthétisé dans l'Hypothèse (H1) suivante.

(H1). *La réponse $Y \in \mathbb{R}$ est définie par le modèle*

$$Y = m(\boldsymbol{X}) + \varepsilon,$$

où le vecteur d'entrée $\boldsymbol{X} = (X^{(1)}, \dots, X^{(p)}) \in [0,1]^p$ admet une densité sur $[0,1]^p$, minorée et majorée par des constantes strictement positives, m est continue, et le bruit ε est sous-gaussien, indépendant de $\boldsymbol{X}$, et centré. Un échantillon $\mathscr{D}_n = \{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)\}$ de n vecteurs aléatoires distribués comme $(\boldsymbol{X}, Y)$ est disponible.

De plus, l'estimateur de la forêt $m_{M,n}(\mathbf{x}, \boldsymbol{\Theta}_M)$ agrège M Θ-arbres aléatoires notés $m_n(\mathbf{x}, \Theta_\ell)$, avec l'aléa de chacun généré par une composante de $\boldsymbol{\Theta}_M = (\Theta_1, \dots, \Theta_M)$: $\Theta_\ell = (\Theta_\ell^{(S)}, \Theta_\ell^{(V)})$, avec $\Theta_\ell^{(S)}$ pour le rééchantillonnage, et $\Theta_\ell^{(V)}$ pour le tirage des variables à chaque noeud.

Dans la définition du MDA de Breiman, le risque quadratique de chaque arbre est estimé à la fois pour l'échantillon "out-of-bag" (OOB) et pour l'échantillon OOB permuté. La différence entre ces deux risques est moyennée sur tous les arbres pour définir le BC-MDA. Plus précisément, pour chaque arbre, nous permutons aléatoirement la j-ième composante de l'ensemble des observations OOB, et notons $\mathbf{X}_{i,\pi_{j\ell}}$ la i-ième observation permutée pour le ℓ-ième arbre. Ainsi, avec $N_{n,\ell} = \sum_{i=1}^n \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$ la taille de l'échantillon OOB du ℓ-ième arbre, le BC-MDA est défini par

$$\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{N_{n,\ell}} \sum_{i=1}^{n} \left[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2 \right] \mathbb{1}_{i \notin \Theta_\ell^{(S)}}.$$

Le TT-MDA est défini de façon similaire, excepté qu'un jeu de données de test est utilisé à la place de l'échantillon OOB, et que le risque quadratique de la forêt est estimé plutôt que celui des arbres dans le cas du BC-MDA. Pour le IK-MDA, l'échantillon OOB est utilisé comme pour le BC-MDA, mais le risque quadratique de l'estimateur OOB de la forêt remplace l'erreur quadratique des arbres.

3 Inconsistance du MDA

Convergence du MDA. A notre connaissance, le Théorème 1 établit le premier résultat de convergence du MDA de Breiman, sous les hypothèses suivantes.

(H2). *L'arbre aléatoire théorique CART construit avec la distribution de* $(\boldsymbol{X}, Y)$ *est consistant, c'est à dire que pour tout* $\boldsymbol{x} \in [0,1]^p$*, presque surement,*

$$\lim_{k\to\infty} \Delta(m, A_k^\star(\boldsymbol{x}, \Theta)) = 0.$$

L'Hypothèse (H2) est toujours vérifiée lorsque la fonction de regression est additive. Ce n'est plus le cas lorsque des termes d'interaction sont présents, car le critère de coupure CART considère les variables une par une. En modifiant légèrement l'algorithme original de Breiman (empêcher les coupures près du bord des cellules), (H2) peut être toujours vérifiée. Enfin, l'Hypothèse (H3) contrôle la complexité de la partition des arbres par rapport à la taille de l'échantillon pour garantir la consistance. Par ailleurs, nous notons $\mathbf{X}_{\pi_j}$ le vecteur $\mathbf{X}$ dont la j-ème composante est remplacée par une copie indépendante de $X^{(j)}$.

(H3). *Le regime asymptotique de* a_n*, la taille du rééchantillonage sans remise, et le nombre de feuilles terminales* t_n *sont tels que* $a_n \leq n-2$*,* $a_n/n < 1-\kappa$ *pour* $\kappa > 0$ *fixé,* $\lim_{n\to\infty} a_n = \infty$*,* $\lim_{n\to\infty} t_n = \infty$*, et* $\lim_{n\to\infty} t_n \frac{(\log(a_n))^9}{a_n} = 0$.

Théorème 1. *Si les Hypothèses (H1), (H2) et (H3) sont vérifiées, alors pour* $M \in \mathbb{N}^\star$ *et* $j \in \{1, \ldots, p\}$*,*

$$(i) \quad \widehat{MDA}_{M,n}^{(TT)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\boldsymbol{X}) - m(\boldsymbol{X}_{\pi_j}))^2]$$
$$(ii) \quad \widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\boldsymbol{X}) - m(\boldsymbol{X}_{\pi_j}))^2].$$

De plus, si $M \underset{n\to\infty}{\longrightarrow} \infty$*, alors*

$$(iii) \quad \widehat{MDA}_{M,n}^{(IK)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\boldsymbol{X}) - \mathbb{E}[m(\boldsymbol{X}_{\pi_j})|\boldsymbol{X}^{(-j)}])^2].$$

Décomposition du MDA. Les limites du MDA peuvent se décomposer en utilisant les indices de Sobol, des mesures bien définies de la contribution de chaque variable d'entrée à la variance de la sortie. Un exemple en dimension deux est donné par la Figure 1. Tout d'abord, nous rappelons la définition formelle de ces indices. L'indice de Sobol total de la variable $X^{(j)}$ (Sobol, 1993) donne la part de variance expliquée de la sortie perdue lorsque $X^{(j)}$ est retirée du modèle, c'est à dire

$$ST^{(j)} = \frac{\mathbb{E}[\mathbb{V}(m(\mathbf{X})|\mathbf{X}^{(-j)})]}{\mathbb{V}(Y)}.$$

L'indice de Sobol total "full" de la variable $X^{(j)}$ (Mara et al., 2015) donne la part de variance expliquée par $X^{(j)}$ en incluant sa contribution due aux interactions et à la dépendance avec les autres entrées, c'est à dire

$$ST^{(j)}_{full} = \frac{\mathbb{E}[\mathbb{V}(m(\mathbf{X}_{\pi_j})|\mathbf{X}^{(-j)})]}{\mathbb{V}(Y)}.$$

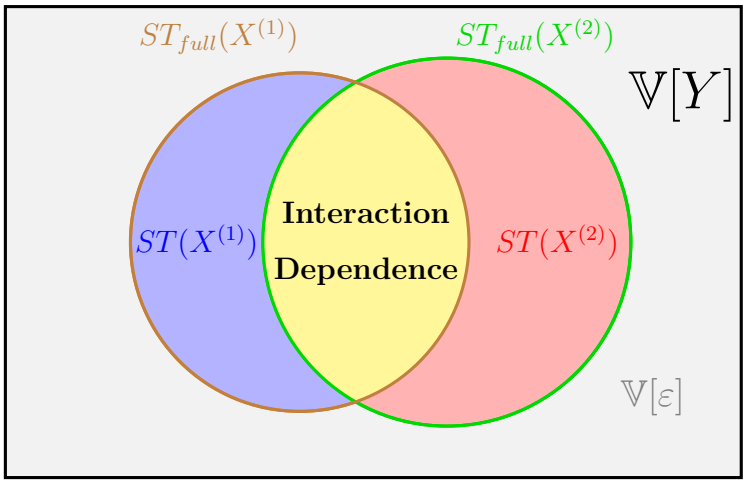


FIGURE 1 – Illustration des indices de Sobol totaux pour $Y = m(X^{(1)}, X^{(2)}) + \varepsilon$.

Ainsi, il est possible de décomposer les limites du MDA à partir des indices de Sobol totaux et du terme $\text{MDA}_3^{\star(j)}$, commenté plus bas et défini par

$$\text{MDA}_3^{\star(j)} = \mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}] - \mathbb{E}[m(\mathbf{X}_{\pi_j})|\mathbf{X}^{(-j)}])^2].$$

Proposition 1. *Si les Hypothèses (H1), (H2) and (H3) sont vérifiées, alors pour* $M \in \mathbb{N}^\star$ *et* $j \in \{1, \ldots, p\}$

$$(i) \quad \widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST^{(j)}_{full} + MDA_3^{\star(j)}$$
$$(ii) \quad \widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST^{(j)}_{full} + MDA_3^{\star(j)}.$$

De plus, si $M \underset{n\to\infty}{\longrightarrow} \infty$, *alors*

$$(iii) \quad \widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + MDA_3^{\star(j)}.$$

Il est essentiel de remarquer que le terme $\text{MDA}_3^{\star(j)}$ n'est pas lié à une mesure d'importance. En particulier, $\text{MDA}_3^{\star(j)}$ est nul lorsque la fonction de régression est additive ou que les entrées sont indépendantes. Dans le cas général, $\text{MDA}_3^{\star(j)}$ peut accroître

	BC-MDA*	$\widehat{\text{BC-MDA}}$	IK-MDA*	$\widehat{\text{IK-MDA}}$	ST*	$\widehat{\text{S-MDA}}$
$\mathbf{X}^{(3)}$	0.47	0.37	0.47	0.43	0.47	0.45
$\mathbf{X}^{(4)}$	0.21	0.10	0.37	0.14	0.10	0.08
$\mathbf{X}^{(5)}$	0.21	0.09	0.37	0.13	0.10	0.08
$\mathbf{X}^{(1)}$	0.64	0.24	1.0	0.29	0.07	0.05
$\mathbf{X}^{(2)}$	0.64	0.24	1.0	0.28	0.07	0.05

Tableau 1 – BC-MDA normalisé, IK-MDA normalisé, et Sobol-MDA.

considérablement la valeur du MDA sans une interprétation claire. Par conséquent, lorsque les variables d'entrées sont dépendantes, le MDA peut conduire à une identification fortement biaisée des variables influentes. Ce phénomène a déjà été observé dans plusieurs études empiriques.

4 Sobol-MDA

Nous proposons l'algorithme Sobol-MDA, une nouvelle mesure d'importance de variables pour les forêts aléatoires qui corrige les défauts du MDA d'origine. Le Sobol-MDA n'est pas basé sur des permutations, mais utilise des projections de la partition des arbres pour éliminer une variable donnée du mécanisme de prédiction. Nous démontrons que le Sobol-MDA estime l'indice de Sobol total de façon consistante, même lorsque les variables d'entrées sont dépendantes, par opposition au MDA de Breiman. Le Tableau 1 fournit des expériences pour une fonction de régression simple et un vecteur d'entrée gaussien corrélé de dimension 5. Seul le Sobol-MDA classe les variables dans l'ordre approprié de l'indice de Sobol total théorique.

Références

L. Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001.

B. Gregorutti. *Random forests and variable selection : analysis of the flight data recorders for aviation safety.* Theses, Université Pierre et Marie Curie - Paris VI, March 2015.

T. A Mara, S. Tarantola, and P. Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72 :173–183, 2015.

I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1 :407–414, 1993.

Accélération d'Anderson pour la descente par coordonnée

Quentin Bertrand [1] & Mathurin Massias [2]

[1] *Université Paris-Saclay, Inria, CEA, Palaiseau, France*
[2] *MaLGa, DIBRIS, University of Genova, Italy*

Résumé. Pour de nombreux problèmes d'apprentissage automatique, la descente par coordonnée permet d'obtenir des performances nettement supérieures à la descente de gradient classique. Accélérer la descente par coordonnée en pratique n'est pas chose facile : les versions inertielles de la descente par coordonnée sont théoriquement accélérées, mais ne conduisent pas toujours à une accélération de la convergence en pratique. Nous proposons une version accélérée de la descente par coordonnée via l'extrapolation non linéaire, montrant des bénéfices pratiques clairs. Des expériences sur les moindres carrés, le Lasso, l'elastic net et la régression logistique montrent un gain pratique significatif.

Mots-clés. Optimisation, optimisation non-lisse, descente par coordonnée, accélération, accélération d'Anderson

Abstract. On multiple Machine Learning problems, coordinate descent achieves performance significantly superior to full-gradient methods. Speeding up coordinate descent in practice is not easy: inertially accelerated versions of coordinate descent are theoretically accelerated, but might not always lead to practical speed-ups. We propose an accelerated version of coordinate descent using extrapolation, showing considerable speed up in practice, compared to inertial accelerated coordinate descent and extrapolated (proximal) gradient descent. Experiments on least squares, Lasso, elastic net and logistic regression validate the approach.

Keywords. Optimization, nonsmooth optimization, coordinate descent, acceleration, Anderson acceleration

1 Introduction

La descente de gradient est la pierre angulaire de l'optimization convexe moderne [11]. Pour les problèmes composites, la descente de gradient proximale est souvent un algorithme de choix. Pour ces deux algorithmes, l'accélération inertielle permet d'obtenir un taux de convergence optimal [10, 2]. La descente par coordonnée est une variante de la descente de gradient, qui met à jour les itérés une coordonnée à la fois [17]. La descente par coordonnée proximale a été appliquée avec succès à de nombreux problèmes

d'apprentissage automatique, en particulier au Lasso [16], à l'elastic net [19] ou à la régression logistique parcimonieuse [13]. Sur le plan théorique, les versions accélérées inertielles de la descente par coordonnée [12, 5] ont des taux de convergence accélérés.

Pour obtenir des algorithmes accélérés, l'accélération d'Anderson [1] est une alternative à l'inertie, qui exploite la structure des itérés. Cette procédure est connue depuis longtemps, sous différents noms [18, 4]. L'accélération d'Anderson a un taux accéléré sur les fonctions quadratiques [7], mais les garanties théoriques dans le cas non quadratique sont plus faibles [15]. L'accélération d'Anderson a été adaptée à de nombreux algorithmes tels que Douglas-Rachford [6], ADMM [14] ou la descente de gradient proximale [8]. Parmi les principaux avantages, la version pratique de l'accélération d'Anderson est efficace en mémoire, facile à implementer, ne requiert pas de recherche linéaire, a un faible coût par itération et ne nécessite pas de connaissance de la constante de forte convexe. Enfin, elle introduit un seul paramètre, qui souvent ne nécessite pas de calibration.

Dans ce travail:

- Nous proposons un schéma d'accélération d'Anderson pour la descente par coordonnée qui, comme visible sur la Figure 2, surpasse le gradient inertiel et extrapolé descente, ainsi que la descente par coordonnée.
- L'accélération est obtenue même si la matrice d'itération n'est pas symétrique, ce qui est une difficulté notable dans l'analyse théorique de l'extrapolation d'Anderson.
- Nous soulignons empiriquement que la technique l'accélération peut se généraliser dans le cas non quadratique (Algorithm 1) et peut significativement améliorer les algorithmes de descente par coordonnée proximales, qui sont à l'état de l'art sur les problèmes considérés.

Une version étendue de ce travail a été publiée dans [3].

2 Algorithme

L'extrapolation d'Anderson accélère la convergence de suites suivant des itérations linéaires de point-fixe, c'est-à-dire:

$$x^{(k+1)} = Tx^{(k)} + b \ , \tag{1}$$

où la matrice d'itération $T \in \mathbb{R}^{p\times p}$ a un rayon spectral $\rho(T) < 1$. La variante hors-ligne de l'extrapolatin d'Anderson construit, à l'itération k, un point fixe approché sous la forme d'une combinaison affine des k premiers itérés: $x^{(k)}_{\text{e-off}} = \sum_1^k c_i^{(k)} x^{(i-1)}$, et les coefficients $c^{(k)} \in \mathbb{R}^k$ sont obtenus comme:

$$\begin{aligned} c^{(k)} &= \underset{\sum_1^k c_i=1}{\arg\min} \|\textstyle\sum_1^k c_i x^{(i-1)} - T\sum_1^k c_i x^{(i-1)} - b\|^2 \\ &= \underset{\sum_1^k c_i=1}{\arg\min} \|\textstyle\sum_1^k c_i \big(x^{(i)} - x^{(i-1)}\big)\|^2 \\ &= (U^\top U)^{-1}\mathbf{1}_k / \mathbf{1}_k^\top (U^\top U)^{-1}\mathbf{1}_k \ , \end{aligned} \tag{2}$$

où $U = [x^{(1)} - x^{(0)}, \dots, x^{(k)} - x^{(k-1)}] \in \mathbb{R}^{p\times k}$ (et l'objective se réécrit donc $\|Uc\|^2$). En pratique, comme $x^{(k)}$ est déjà calculé lorsque $c^{(k)}$ est calculé, on utilise $x_{\mathrm{e}}^{(k)} = \sum_1^k c_i^{(k)} x^{(i)}$ au lieu de $\sum_1^k c_i^{(k)} x^{(i-1)}$. Une justification pour l'introduction des coefficients $c^{(k)}$ est détaillée dans la proposition 6 de [9]. Pour l'extrapolation en ligne, au lieu de considérer les k premiers itérés, on utilise seulement une combinaison des K derniers itérés (où $K = 5$ en pratique). C'est cette variante qui est utilisée en pratique car il est moins coûteux de calculer les coefficients c.

Nous appliquons cette technique d'accélération (Algorithm 1) à des problèmes de la forme:

$$\min_{x\in\mathbb{R}^p} f(Ax) + \lambda g(x) := f(Ax) + \lambda \sum_{j=1}^{p} g_j(x_j) \ , \tag{3}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ est convexe et lisse et les g_j sont convexes, propres et fermées.

Dans le cas des moindres carrés non pénalisés ($f = \frac{1}{2}\|\cdot - y\|^2$, $g_j = 0$), les itérés de la descente par coordonnée (pris à la fin de chaque époque, c'est-à-dire après mise à jour des p coordonnées) possèdent une structure d'itération linéaire:

$$x^{(k+1)} = T^{\mathrm{CD}} x^{(k)} + b^{\mathrm{CD}} \ , \tag{4}$$

avec $T^{\mathrm{CD}} = \left(\mathrm{Id}_p - \frac{e_p e_p^\top}{H_{pp}} H\right) \dots \left(\mathrm{Id}_p - \frac{e_1 e_1^\top}{H_{11}} H\right)$ et $H = A^\top A$. Cette matrice n'est pas symétrique, ce qui est nécessaire pour l'obtention de guaranties théoriques pour l'accélération d'Anderson. Sur le plan théorique, nous proposons d'effectuer les mise à jour des coordonnées de 1 à p, suivi d'une passe inversée, de p à 1. La matrice d'itération obtenue n'est alors pas symétrique, mais elle s'écrit $T^{\text{CD-sym}} \triangleq H^{-1/2} S H^{1/2}$, avec

$$\begin{aligned} S = &\left(\mathrm{Id}_p - H^{1/2}\tfrac{e_1 e_1^\top}{H_{11}} H^{\frac{1}{2}}\right) \times \dots \times \left(\mathrm{Id}_p - H^{\frac{1}{2}}\tfrac{e_p e_p^\top}{H_{pp}} H^{\frac{1}{2}}\right) \\ &\times \left(\mathrm{Id}_p - H^{\frac{1}{2}}\tfrac{e_p e_p^\top}{H_{pp}} H^{\frac{1}{2}}\right) \times \dots \times \left(\mathrm{Id}_p - H^{\frac{1}{2}}\tfrac{e_1 e_1^\top}{H_{11}} H^{\frac{1}{2}}\right) \ . \end{aligned} \tag{5}$$

S est symétrique, et S et T (qui a les mêmes valeurs propres que S) sont diagonalisables avec spectre réel. Nous appelons ces itérations "pseudo-symétriques" et montrons qu'elles permettent de préserver les garanties de l'accélération d'Anderson.

Proposition 1 (Pseudosymétrique, $T = H^{-1/2} S H^{1/2}$) *Soit T la matrice d'itération de la descente par coordonnée pseudo-symétrique: $T = H^{-1/2} S H^{1/2}$, avec S positive semi-définie (5). Soit x^* la limite de la suite $(x^{(k)})$. Soit $\zeta = (1 - \sqrt{1-\rho})/(1 + \sqrt{1-\rho})$, soit $B = (T - \mathrm{Id})^2$ et $\kappa(H)$ le conditionnement de H. Alors $\rho = \rho(T) = \rho(S) < 1$ et les itérés de l'extrapolation d'Anderson hors-ligne satisfont:*

$$\|x_{\text{e-off}}^{(k)} - x^*\|_B \le \sqrt{\kappa(H)} \tfrac{2\zeta^{k-1}}{1+\zeta^{2(k-1)}} \|x^{(0)} - x^*\|_B \ , \tag{6}$$

en donc ceux de l'extrapolation en ligne satisfont:

$$\|x_{\text{e-on}}^{(k)} - x^*\|_B \le \left(\sqrt{\kappa(H)} \tfrac{2\zeta^{K-1}}{1+\zeta^{2(K-1)}}\right)^{k/K} \|x^{(0)} - x^*\|_B \ . \tag{7}$$

Algorithm 1 Accélération d'Anderson pour la descente par coordonnée

init: $x^{(0)} \in \mathbb{R}^p, L_1, \dots, L_p > 0$
for $k = 1, \dots$ **do**
  $x = x^{(k-1)}$
  for $j = 1, \dots p$ **do**
    $\tilde{x}_j = x_j$
    $x_j = \text{prox}_{\frac{\lambda}{L_j} g_j}(x_j - A_{:j}^\top \nabla f(Ax)/L_j)$
    $Ax \mathrel{+}= (x_j - \tilde{x}_j) A_{:j}$
  $x^{(k)} = x$ // itérés classiques $\mathcal{O}(np)$
  if $k = 0 \mod K$ **then** // extrapolation, $\mathcal{O}(K^3 + pK^2)$
    $U = [x^{(k-K+1)} - x^{(k-K)}, \dots, x^{(k)} - x^{(k-1)}]$
    $c = (U^\top U)^{-1} \mathbf{1}_K / \mathbf{1}_K^\top (U^\top U)^{-1} \mathbf{1}_K \in \mathbb{R}^K$
    $x_e = \sum_{i=1}^K c_i x^{(k-K+i)}$
    if $f(Ax_e) + \lambda g(x_e) \leq f(x^{(k)}) + \lambda g(x^{(k)})$ **then**
      $x^{(k)} = x_e$
return $x^{(k)}$

3 Expériences

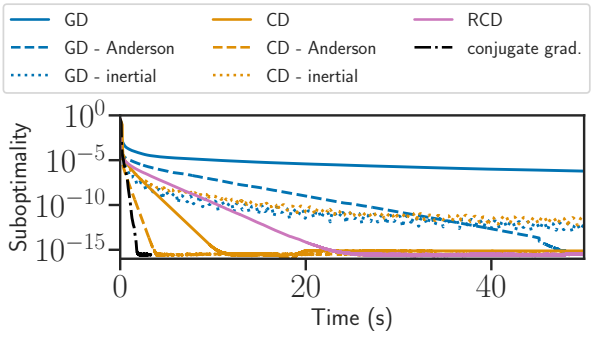


Figure 1: Performance de la descente de gradient (GD), et de la descente par coordonnée (CD) avec leurs versions accélérées. RCD: descente par coordonnée aléatoire.

Sur la Figure 1, on peut observer la supériorité de la descente par coordonnée avec extrapolation d'Anderson sur d'autres méthodes du premier ordre, éventuellement accélérée.
Dans ce qui suit, nous montrons que l'extrapolation d'Anderson appliquée à la descente par coordonnée proximale surpasse les autres algorithmes du premier ordre pour la résolution du Lasso (Figure 2):

$$x^* = \arg\min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1 \quad . \tag{8}$$

Pour le Lasso, nous paramétrons λ comme une fraction de $\lambda_{\max} = \left\|A^\top y\right\|_\infty$, plus petit paramètre de régularisation pour lequel $x^* = 0$. La Figure 2 montre la supériorité de la descente par coordonnée proximale sur la descente de gradient proximale pour les problèmes de Lasso sur des jeux de données réels, ainsi que les avantages de l'extrapolation pour la descente par coordonnée. Elle montre que l'extrapolation d'Anderson peut conduire à un gain de performance significatif. En particulier Figure 2 montre que sans

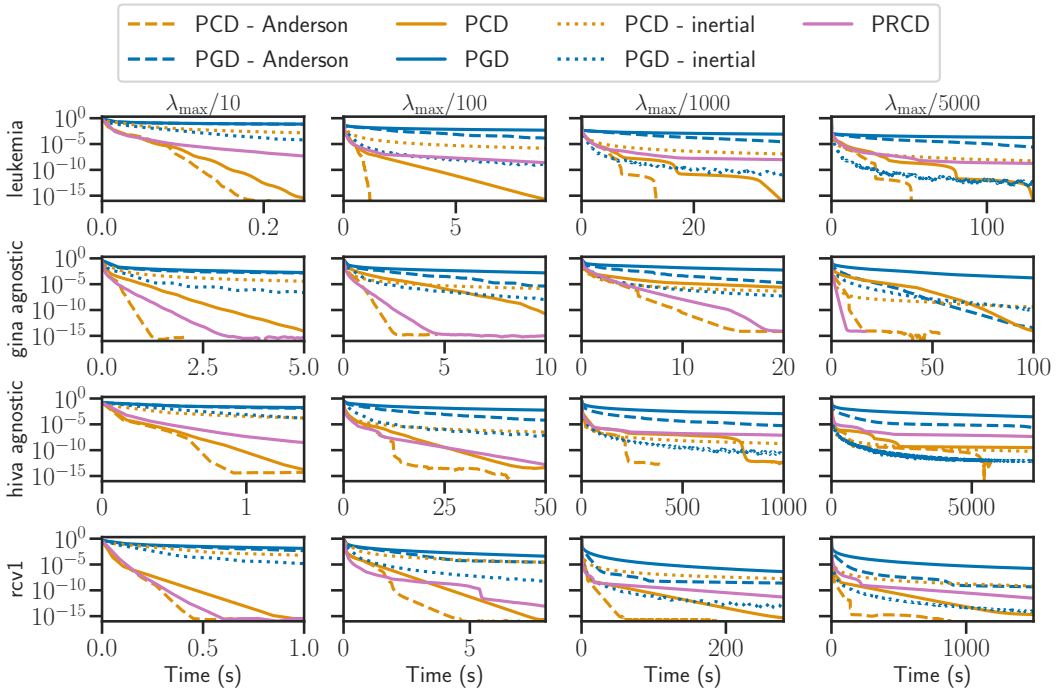


Figure 2: Sous optimalité en fonction du tempspour le Lasso pour différents jeux de données et valeurs de λ.

redémarrage (*restart*), la descente par coordonnée inertielle peut ralentir la convergence en pratique, malgré son taux accéléré en théorie. Notons que plus la valeur de λ est petite, plus l'optimisation est difficile: lorsque λ diminue, atteindre une sous-optimalité fixé prend plus de temps. Plus le paramètre de régularisation λ est petit (*i.e.,* plus le problème est difficile), plus l'extrapolation d'Anderson est efficace. Les noms et les détails des algorithmes comparés peuvent être trouvés en section 3 de [3].

References

[1] D. G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560, 1965.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[3] Q. Bertrand and M. Massias. Anderson acceleration of coordinate descent. *AISTATS*, 2021.

[4] R. P. Eddy. Extrapolating to the limit of a vector sequence. In *Information linkage between applied mathematics and industry*, pages 387–396. Elsevier, 1979.

[5] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[6] A. Fu, J. Zhang, and S. Boyd. Anderson accelerated Douglas-Rachford splitting. *arXiv preprint arXiv:1908.11482*, 2019.

[7] G. H. Golub and R. S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods. *Numerische Mathematik*, 3(1):147–156, 1961.

[8] V. V. Mai and M. Johansson. Anderson acceleration of proximal gradient methods. In *ICML*. 2019.

[9] M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. Dual extrapolation for sparse generalized linear models. *J. Mach. Learn. Res.*, 2020.

[10] Y. Nesterov. A method for solving a convex programming problem with rate of convergence $O(1/k^2)$. *Soviet Math. Doklady*, 269(3):543–547, 1983.

[11] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic publishers, Boston, MA, 2004.

[12] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[13] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *ICML*, page 78, 2004.

[14] C. Poon and J. Liang. Trajectory of alternating direction method of multipliers and adaptive acceleration. In *NeurIPS*, pages 7357–7365, 2019.

[15] D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

[17] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

[18] P. Wynn. Acceleration techniques for iterated vector and matrix problems. *Mathematics of Computation*, 16(79):301–322, 1962.

[19] H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Nonlinear Functional-Output Regression: A Dictionary Approach

Dimitri Bouche [1] & Marianne Clausel [2] & François Roueff [1] & Florence d'Alché-Buc [1]

[1] *LTCI, Télécom Paris, Institut Polytechnique de Paris; first.last@telecom-paris.fr*
[2] *Université de Lorraine, CNRS, IECL; marianne.clausel@univ-lorraine.fr*

Résumé. Dans le contexte de la regression à valeur fonctionnelle, nous proposons d'apprendre à prédire directement des coefficients de représentation sur un dictionnaire en employant toutefois une fonction de perte fonctionnelle. Ainsi, un dictionnaire non-orthogonal peut être choisi; il peut donc être appris. En cela l'approche est plus flexible que celles employant une perte vectorielle sur les coefficients. Nous étudions en détails le cas où la fonction prédisant les coefficients réside dans un espace de Hilbert à noyau reproduisant de fonctions à valeurs vectorielles. Pour la perte quadratique nous introduisons deux estimateurs en forme close, le premier pour des fonctions totalement observées et le second pour des fonctions partiellement observées. Les deux sont appuyés par une analyse de leur excès de risque. Enfin, nous démontrons sur deux jeux de données réels que notre approche constitue un bon compromis entre temps de calcul et précision.

Mots-clés. Données fonctionnelles, Noyaux reproduisant, Dictionnaires.

Abstract. To address functional-output regression, we introduce *projection learning* (PL), a novel dictionary-based approach that learns to predict a projection of the output function on a dictionary while minimizing a functional loss. PL makes it possible to use non orthogonal dictionaries and can then be combined with dictionary learning. It is thus much more flexible than expansion-based approaches relying on vectorial losses. Using reproducing kernel Hilbert spaces of vector-valued functions, this general method is instantiated as *kernel-based projection learning* (KPL). For the functional square loss, we propose two closed-form estimators, one for fully observed output functions and the other for partially observed ones. Both are backed theoretically by an excess risk analysis. Eventually, a study on two real datasets show that our approach enjoys a good trade-off between computational cost and precision.

Keywords. Functional data, Reproducing kernel Hilbert spaces, Dictionaries.

1 Notations

We assimilate the spaces $(\mathbb{R}^d)^n$ and $\mathbb{R}^{d\times n}$. The concatenation of vectors $(u_i)_{i=1}^n \in \mathbb{R}^{d\times n}$ is denoted $\mathsf{vec}((u_i)_{i=1}^n) \in \mathbb{R}^{dn}$. For $n \in \mathbb{N}^*$, we use the shorthand $[n]$ for the set $\{1,\dots,n\}$. We denote by $\mathcal{F}(\mathcal{X},\mathcal{Y})$ the space of functions from $\mathcal{X}$ to $\mathcal{Y}$. For two Hilbert spaces $\mathcal{U}$ and

$\mathcal{Y}$, $\mathcal{L}(\mathcal{U},\mathcal{Y})$ is the set of bounded linear operators from $\mathcal{U}$ to $\mathcal{Y}$ and $\mathcal{L}(\mathcal{U}) := \mathcal{L}(\mathcal{U},\mathcal{U})$. The adjoint of a linear operator A is denoted $\mathsf{A}^{\#}$. For $\mathcal{U} = \mathbb{R}^d$, we introduce $\mathsf{A}_{(n)} \in \mathcal{L}(\mathbb{R}^{dn}, \mathcal{Y}^n)$ as $\mathsf{A}_{(n)} : \mathsf{vec}((u_i)_{i=1}^n) \longmapsto (\mathsf{A}u_1, ..., \mathsf{A}u_n)$ and $\mathsf{A}_{\mathsf{mat},(n)} \in \mathcal{L}(\mathbb{R}^{d\times n}, \mathcal{Y}^n)$ as $\mathsf{A}_{\mathsf{mat},(n)} : (u_i)_{i=1}^n \longmapsto (\mathsf{A}u_1, ..., \mathsf{A}u_n)$. For $\mathsf{B} \in \mathbb{R}^{p\times q}, \mathsf{C} \in \mathbb{R}^{d\times n}$, $\mathsf{B} \otimes \mathsf{C} \in \mathbb{R}^{pd\times qn}$ denotes the Kronecker product. Finally $\mathsf{L}^2(\Theta)$ stands for the Hilbert space of real-valued square integrable functions on a given compact subset $\Theta \subset \mathbb{R}^q$; without loss of generality we suppose that $|\Theta| := \int_\Theta 1 \mathrm{d}\theta = 1$.

2 Projection learning (PL)

2.1 Functional output regression (FOR)

Let $\mathcal{X}$ be a measurable space and (X, Y) be a couple of random variables on $\mathcal{Z} := \mathcal{X} \times \mathsf{L}^2(\Theta)$ with joint distribution ρ. We define a functional loss ℓ as a real-valued function over $\mathsf{L}^2(\Theta) \times \mathsf{L}^2(\Theta) \to \mathbb{R}$. An integral of a ground loss $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ can for instance be used:

$$\ell(y_0, y_1) = \int_\Theta l(y_0(\theta), y_1(\theta)) \mathrm{d}\theta. \tag{1}$$

Specifically, if $l(y_0(\theta), y_1(\theta)) = (y_0(\theta) - y_1(\theta))^2$, we obtain $\ell_2(y_0, y_1) := \|y_0 - y_1\|^2_{\mathsf{L}^2(\Theta)}$. Given a loss ℓ we define the risk of a regressor f as $\mathcal{R}(f) := \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\rho}[\ell(\mathsf{Y}, f(\mathsf{X}))]$. For an hypothesis class $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathsf{L}^2(\Theta))$ we then want to find a minimizer of this risk:

$$\arg\min_{f\in\mathcal{G}} \mathcal{R}(f). \tag{2}$$

However, we have access to ρ only through an observed sample. We study two sampling settings. **(i)** The output functions are *fully observed*. Our sample consists of $n \in \mathbb{N}$ i.i.d. realizations $\mathbf{z} := (x_i, y_i)_{i=1}^n$ drawn from ρ (so-called *dense* setting described in functional data analysis (FDA)). **(ii)** In the *partially observed* setting (*sparse* setting in FDA) they are observed on discrete grids. In this work, we suppose that we observe each y_i on a random sample of locations, $\theta_i := (\theta_{ip})_{p=1}^{m_i} \in \Theta^{m_i}$, drawn i.i.d.from a uniform distribution $\widetilde{\mathbf{z}} := (x_i, (\theta_i, \widetilde{y}_i))_{i=1}^n$, where for all $i \in [n]$, $\theta_i \in \Theta^{m_i}$, $\widetilde{y}_i \in \mathbb{R}^{m_i}$ with $m_i \in \mathbb{N}^*$ the number of observations available for the i-th function, and for all $p \in [m_i]$, $\theta_{ip} \in \Theta$ and $\widetilde{y}_{ip} \in \mathbb{R}$.

2.2 Approximated FOR

To tackle Problem (2), we propose to learn to predict expansion coefficients on a dictionary of functions $\phi := (\phi_l)_{l=1}^d \in \mathsf{L}^2(\Theta)^d$ with $d \in \mathbb{N}^*$. We then introduce the linear operator:

Definition 2.1. (Projection operator) For a dictionary ϕ, the associated projection operator Φ is defined by $\Phi : \quad u \in \mathbb{R}^d \longmapsto \sum_{l=1}^d u_l \phi_l \in \mathsf{L}^2(\Theta)$.

Lemma 2.1. *The adjoint of* Φ *is given by* $\Phi^{\#} : \ g \in \mathsf{L}^2(\Theta) \longmapsto (\langle \phi_l, g \rangle_{\mathsf{L}^2(\Theta)})_{l=1}^d \in \mathbb{R}^d$. *Thus we have* $\Phi^{\#}\Phi = (\langle \phi_l, \phi_s \rangle_{\mathsf{L}^2(\Theta)})_{l,s=1}^d$.

The core idea of PL is to define a simpler model $f(x) = \Phi h(x)$ in Problem (2), where $h : \mathcal{X} \longmapsto \mathbb{R}^d$ is a vector-valued function. This yields the problem

$$\arg\min_{h \in \mathcal{H}} \mathcal{R}(\Phi \circ h). \tag{3}$$

In the fully observed setting, we can minimize over $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ the empirical counterpart of the true risk based on $\mathbf{z}$, $\widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i))$ with some additional penalty $\Omega_{\mathcal{H}} : \mathcal{H} \longrightarrow \mathbb{R}$ of intensity $\lambda > 0$ to control the model's complexity: $\min_{h \in \mathcal{H}} \widehat{\mathcal{R}}(\Phi \circ h, \mathbf{z}) + \lambda \Omega_{\mathcal{H}}(h)$. To tackle the partially observed setting, we exploit specific properties of the learning algorithms proposed in Section 3.

2.3 Dictionaries

In solving Problem (3) we restrict the predictions of our model to Span(ϕ). As a result ϕ must be chosen so that the functions $(y_i)_{i=1}^n$ can be approximated accurately by elements from Span(ϕ). To achieve this, several strategies are possible.
Orthonormal and Riesz bases. Orthogonal bases such as Fourier bases or wavelets bases, as well as Riesz bases such as splines, are known to provide good approximations.
Families of random functions, such as random Fourier features (RFFs, Rahimi and Recht, 2008) can enjoy good approximation properties as well. Through the choice of such family, we approximate the output functions in a space that is dense in a RKHS.
Learnt dictionaries. When the output functions are too complex, selecting a dictionary can however be difficult. The choice of a family may not be evident and it may take too many atoms (functions) to reach a satisfying approximation precision. Functional principal component analysis (FPCA; Ramsay and Silverman, 2005) addresses the first issue, but may not adress the second. By opposition, dictionary learning (DL) is known to better solve both problems (Mairal et al., 2009).

3 Vector-valued RKHSs (Vv-RKHS) instantiation

3.1 Vv-RKHSs and representer theorem

Let $\mathsf{K} : \mathcal{X} \times \mathcal{X} \longmapsto \mathcal{L}(\mathbb{R}^d)$ be operator-valued kernel and $\mathcal{H}_{\mathsf{K}} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$ its associated vv-RKHS. For $x \in \mathcal{X}$, we define $\mathsf{K}_x \in \mathcal{L}(\mathbb{R}^d, \mathcal{H}_{\mathsf{K}})$ as $\mathsf{K}_x : u \longmapsto \mathsf{K}_x u$, with $\mathsf{K}_x u : x' \longmapsto \mathsf{K}(x', x)u$. We consider $\mathcal{H} = \mathcal{H}_{\mathsf{K}}$ as vector-valued hypothesis class. Setting the regularization as $\Omega_{\mathcal{H}_{\mathsf{K}}}(h) := \|h\|^2_{\mathcal{H}_{\mathsf{K}}}$ yields the following instantiation of PL with vv-RKHS:

$$\arg\min_{h \in \mathcal{H}_{\mathsf{K}}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi h(x_i)) + \lambda \|h\|^2_{\mathcal{H}_{\mathsf{K}}}. \tag{4}$$

Proposition 3.1. ***(Representer theorem)*** *For ℓ continuous and convex with respect to its second argument, Problem (4) admits a unique minimizer $h^\lambda_{\mathbf{z}}$. Moreover there exists $\alpha \in \mathbb{R}^{d\times n}$ such that $h^\lambda_{\mathbf{z}} = \sum_{j=1}^n \mathsf{K}_{x_j}\alpha_j$.*

3.2 Ridge solution with square loss

Problem (4) can be rewritten as $\min_{\alpha\in\mathbb{R}^{d\times n}} \frac{1}{n}\left\|\mathbf{y} - \Phi_{(n)}\mathbf{K}\mathsf{vec}(\alpha)\right\|^2_{\mathsf{L}^2(\Theta)^n} + \lambda\langle \mathsf{vec}(\alpha), \mathbf{K}\mathsf{vec}(\alpha)\rangle_{\mathbb{R}^{dn}}$, where $\mathbf{y} := (y_i)_{i=1}^n \in \mathsf{L}^2(\Theta)^n$, the kernel matrix is defined block-wise as $\mathbf{K} := [\mathsf{K}(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{dn\times dn}$; and vec and $\Phi_{(n)}$ are introduced in Section 1.

Proposition 3.2. ***(Ridge solution)*** *If $\mathbf{K}$ is full rank, the minimum in the above stated Problem is achieved by the unique $\alpha^* \in \mathbb{R}^{d\times n}$ verifying $\left((\Phi^\#\Phi)_{(n)}\mathbf{K} + n\lambda\mathbf{I}\right)\mathsf{vec}(\alpha^*) = \Phi^\#_{(n)}\mathbf{y}$. We define the ridge estimator as $h^\lambda_{\mathbf{z}} := \sum_{j=1}^n \mathsf{K}_{x_j}\alpha^*_j$.*

$(\Phi^\#\Phi)_{(n)}$ is a block diagonal matrix with the Gram matrix $\Phi^\#\Phi$ repeated on its diagonal. To handle partially observed output functions, we remark that in the above linear system, the output functions only appear through $(\Phi_{(n)})^\#\mathbf{y} = \mathsf{vec}((\Phi^\# y_i)_{i=1}^n) \in \mathbb{R}^{dn}$ with for $i \in [n]$, $\Phi^\# y_i = \left(\langle y_i, \phi_l\rangle_{\mathsf{L}^2(\Theta)}\right)_{l=1}^d$.

Definition 3.1. (**Plug-in ridge estimator.**) For all $l \in [d]$ and $i \in [n]$, let $\widetilde{\nu}_{il} := \frac{1}{m_i}\sum_{p=1}^{m_i} \widetilde{y}_{ip}\phi_l(\theta_{ip})$ be the entries of $\widetilde{\nu} \in \mathbb{R}^{d\times n}$. Let $\widetilde{\alpha}^* \in \mathbb{R}^{d\times n}$ be such that $\mathsf{vec}(\widetilde{\alpha}^*) = \left((\Phi^\#\Phi)_{(n)}\mathbf{K} + n\lambda\mathbf{I}\right)^{-1}\mathsf{vec}(\widetilde{\nu})$. We then define the plug-in ridge estimator as $\widetilde{h}^\lambda_{\mathbf{z}} := \sum_{j=1}^n \mathsf{K}_{x_j}\widetilde{\alpha}^*_j$.

Fast algorithm for plug-in ridge estimator. For a separable kernel $\mathsf{K} = k\mathsf{B}$, the matrix $\mathbf{K}$ can be rewritten as $\mathbf{K} = \mathsf{K}_\mathcal{X} \otimes \mathsf{B}$ with $\mathsf{K}_\mathcal{X} := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$. Solving the linear system in Proposition 3.2 has time complexity $\mathcal{O}(n^3d^3)$. However, $(\Phi_{(n)})^\#\Phi_{(n)} = \mathsf{I}\otimes(\Phi^\#\Phi)$, thus $(\Phi_{(n)})^\#\Phi_{(n)}\mathbf{K} = (\mathsf{I}\otimes(\Phi^\#\Phi))(\mathsf{K}_\mathcal{X}\otimes\mathsf{B})$. Using the mixed product property we must solve $(\mathsf{K}_\mathcal{X}\otimes((\Phi^\#\Phi)\mathsf{B}) + n\lambda\mathbf{I})\mathsf{vec}(\alpha) = \mathsf{vec}(\widetilde{\nu})$. The complexity can be reduced essentially to $\mathcal{O}(n^3 + d^3)$ using either a discrete time Sylvester equation or performing an eigendecomposition exploiting the Kronecker structure. For other losses, we resort to iterative optimization; we refer to Bouche et al. (2021) for details.

4 Excess risk analysis

For the ridge estimator, we focus on the effect of the number of samples n, and for the plug-in ridge one, we study the influence of n and that of m (number of observations per function). The analysis is based on integral operators (Caponnetto and De Vito, 2007).

Let us suppose that $\mathcal{X}$ is a separable metric space. K is a vector-valued continuous kernel and we suppose $\exists\kappa > 0$ such that for $x \in \mathcal{X}$, $\|\mathsf{K}(x,x)\|_{\mathcal{L}(\mathbb{R}^d)} \leq \kappa$. ϕ is a normed Riesz family in $\mathsf{L}^2(\Theta)$ with upper constant C_ϕ. We also suppose that $\exists h_{\mathcal{H}_\mathsf{K}} \in \mathcal{H}_\mathsf{K}$ such that $h_{\mathcal{H}_\mathsf{K}} = \inf_{h\in\mathcal{H}_\mathsf{K}} \mathcal{R}(\Phi\circ h)$ which implies $\exists R > 0$ such that $\|h_{\mathcal{H}_\mathsf{K}}\|_{\mathcal{H}_\mathsf{K}} \leq R$. Finally, we assume that $\exists L \geq 0$ such that for all $\theta \in \Theta$, $|\mathsf{Y}(\theta)| \leq L$ a. s.

Proposition 4.1. *Let* $0 < \eta < 1$, *take* $\lambda = \lambda_n^*(\eta/2) := 6\kappa C_\phi^2 \frac{\log(4/\eta)\sqrt{d}}{\sqrt{n}}$. *Then with probability at least* $1-\eta$, $\mathcal{R}(\Phi \circ h_{\mathbf{z}}^{\lambda}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \log(4/\eta)$.

We suppose that $\exists M(d) \geq 0$ such that for all $\theta \in \Theta$ and for all $l \in [d]$, $|\phi_l(\theta)| \leq M(d)$.

Proposition 4.2. *Let* $0 < \eta < 1$, *take* $\lambda = \lambda_n^*(\eta/3) := 6\kappa C_\phi^2 \frac{\log(6/\eta)\sqrt{d}}{\sqrt{n}}$. *Then, with probability at least* $1-\eta$, $\mathcal{R}(\Phi \circ \widetilde{h}_{\widetilde{\mathbf{z}}}^{\lambda}) - \mathcal{R}(\Phi \circ h_{\mathcal{H}_K}) \leq \left(\mathcal{O}\left(\frac{\sqrt{n}}{m^2}\right) + \mathcal{O}\left(\frac{1}{m^{3/2}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}m}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \log(6/\eta)$.

If $m \asymp \sqrt{n}$, then this bounds yields consistency for the plug-in ridge estimator.

5 Experiments

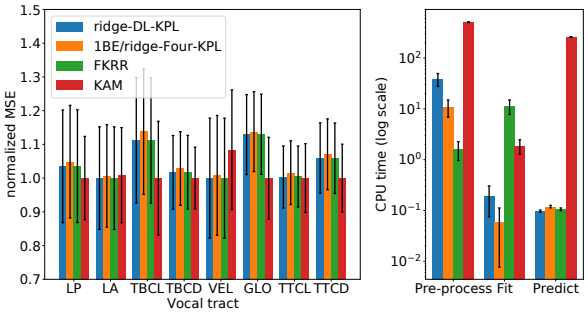


Figure 1: MSEs and CPU times on the speech data

KE	0.231 ± 0.025
3BE	0.227 ± 0.017
KAM	0.222 ± 0.021
FKRR	0.215 ± 0.020
RIDGE-KPL	0.211 ± 0.022
LOGCOSH-KPL	**0.209** ± **0.020**

Table 1: MSEs on the DTI data

We compare KPL with four existing nonlinear FOR methods. Functional kernel ridge regression (FKRR, Kadri et al., 2016), Triple basis estimator (3BE, Oliva et al., 2015) (which we call 1BE when using a kernel for the inputs and only the outputs are projected on a basis), Kernel additive model (KAM, Reimherr et al., 2018) and Kernel Estimator (KE, Ferraty et al., 2011).

We consider two datasets. The first is the diffusion tensor imaging dataset (DTI data) [1], and the second is a synthetic speech inversion dataset introduced by Mitra et al. (2009). The first is a function-to-function problem and the second one is a sound-to-functions problem; from a speech sound we estimate the underlying vocal tract (VT) configuration that produced it (there are eight VTs). We represent the input sounds using 13 mel-frequency cepstral coefficients (MFCCs), so that the problem becomes a vector-valued function-to-function task. The mean square errors for the DTI datasets are displayed in Table 1. The results and computational times for the speech dataset are displayed in Figure 1. *Ridge-DL-KPL* corresponds to the plug-in ridge estimatore with a learnt

[1] This dataset was collected at Johns Hopkins University and the Kennedy-Krieger Institute and is freely available as a part of the *Refund* R package

dictionary and *1BE/ridge-Four-KPL* to the same estimator using a truncated Fourier basis as dictionary. We see that our proposed method enjoys a particularily good computation time/precision trade-off. For more experiences and details on the experimental procedures we refer to (Bouche et al., 2021).

References

D. Bouche, M. Clausel, F. Roueff, and F. d'Alché Buc. Nonlinear functional output regression: A dictionary approach. In *Proceedings of AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, 2021.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, pages 331–368, 2007.

F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu. Kernel regression with functional response. *Electron. J. Statist.*, 5:159–171, 2011.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17:1–54, 2016.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 689–696, 2009.

V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C. Y. Espy-Wilson. From acoustics to vocal tract time functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4500, 2009.

J. Oliva, W. Neiswanger, B. Poczos, E. Xing, H. Trac, S. Ho, and J. Schneider. Fast function to function regression. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38, pages 717–725, 2015.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 20*, pages 1177–1184. Curran Associates, Inc., 2008.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, 2005.

M. Reimherr, B. Sriperumbudur, and B. Taoufik. Optimal prediction for additive function on function regression. *Electronic Journal of Statistics*, 12:4571–4601, 2018.

Making the most of your day: online learning for optimal allocation of time

Etienne Boursier [1] & Tristan Garrec [1,2] & Vianney Perchet [3] & Marco Scarsini [4]

[1] *Université Paris-Saclay, ENS Paris-Saclay, Centre Borelli, Gif-sur-Yvette, France*
eboursie@ens-paris-saclay.fr
[2] *EDF Lab, Gif-sur-Yvette, France*
tristan.garrec@ut-capitole.fr
[3] *CREST, ENSAE Paris, Palaiseau & CRITEO AI Lab, Paris, France*
vianney.perchet@normalesup.org
[4] *Université LUISS, Rome, Italie*
marco.scarsini@luiss.it

Résumé. Nous étudions l'apprentissage séquentiel d'une allocation optimale lorsque la ressource à attribuer est le temps. Notre modèle comprend par exemple les applications suivantes : un chauffeur décidant de ses trajets quotidiens, un propriétaire louant son bien, etc. Suivant notre motivation initiale, un chauffeur reçoit de manière séquentielle des propositions de trajet selon un processus de Poisson. Il peut alors soit accepter, soit refuser le trajet. S'il accepte, le chauffeur est occupé pendant la durée du trajet et obtient un revenu dépendant de cette durée. Sinon, le chauffeur reste en attente jusqu'à recevoir une nouvelle proposition de trajet. Nous étudions le regret encouru par le chauffeur, d'abord s'il connaît la fonction de revenu mais ignore la distribution des durées de trajet, puis s'il ne connaît pas la fonction de revenu non plus. Ce cadre naturel partage des similarités avec les bandits contextuels à un bras, mais avec la différence essentielle qu'ici, le revenu standard associé à un contexte dépend de toute la distribution des contextes.

Mots-clés. Service de transport, apprentissage séquentiel, allocation temporelle, bandits contextuels

Abstract. We study online learning for optimal allocation when the resource to be allocated is time. Examples of possible applications include a driver filling a day with rides, a landlord renting an estate, etc. Following our initial motivation, a driver receives ride proposals sequentially according to a Poisson process and can either accept or reject a proposed ride. If she accepts the proposal, she is busy for the duration of the ride and obtains a reward that depends on the ride duration. If she rejects it, she remains on hold until a new ride proposal arrives. We study the regret incurred by the driver first when she knows her reward function but does not know the distribution of the ride duration, and then when she does not know her reward function, either. Faster rates are finally obtained by adding structural assumptions on the distribution of rides or on the reward function. This natural setting bears similarities with contextual (one-armed) bandits, but with the crucial difference that the normalized reward associated to a context depends on the whole distribution of contexts.

Keywords. Ride hailing, online learning, time allocation, contextual bandits

1 Motivation

A driver filling her shift with rides, a landlord renting an estate short-term, an independent deliveryman, a single server that can make computations online, etc. all face the same trade-off. There is a unique resource that can be allocated to some tasks/clients for some duration. The main constraint is that, once it is allocated, the resource becomes unavailable for the whole duration. As a consequence, if a "better" request arrived during this time, it could not be accepted and would be lost. Allocating the resource for some duration has some cost but generates some rewards – possibly both unknown and random. For instance, an estate must be cleaned up after each rental, thus generating some fixed costs; on the other hand, guests might break something, which explains randomness. Similarly, the driver net revenue might depend on the traffic jam and/or the weather (through gas consumption). Concerning duration, the shorter the request the better (if the net reward is the same). Indeed, the resource could be allocated twice in the same amount of time.

The ideal request would therefore be of short duration and large reward; this maximizes the revenue per time. A possible policy could be to wait for this kind of request, denying the other ones (too long and/or less profitable). On the other hand, such a request could be very rare. So it might be more rewarding in the long run to accept any request, at the risk of "missing" the ideal one.

There are clear trade-offs that arise. The first one is between a greedy policy that accepts only the highest profitable requests – at the risk of staying idle quite often – and a safe policy that accepts every request – but unfortunately also the non-profitable ones. The second trade-off concerns the learning phase; indeed, at first and because of the randomness, the actual net reward of a request is unknown and must be learned on the fly. The safe policy will gather a lot of information (possibly at a high cost) while the greedy one might lose some possible valuable information for the long run (in trying to optimize the short term revenue).

We adopt the perspective of a driver seeking to optimize the income obtained during a shift. The driver receives ride proposals sequentially, following a Poisson process. When a ride is proposed, the driver observes its expected duration and can then either accept or reject it. If she accepts it, she cannot receive any new proposals for the whole duration of the ride. At the end of the ride she observes her reward, which is a function of the duration of the ride. If, on the contrary, she rejects the ride, she remains on hold until she receives a new ride proposal.

Driver policies are evaluated in terms of their expected regret, which is the difference between the cumulative rewards obtained over a shift of duration T under the optimal

policy and under the implemented driver policy (as usual the total length could also be random or unknown [Degenne and Perchet, 2016]). In this setting, the "optimal" policy is within the class of policies that accept – or not – rides whose length belongs to some given acceptance set (say, larger than some threshold, or in some specific Borel subset, depending on the regularity of the reward function).

2 Model

We consider a driver who receives ride proposals sequentially and decides whether to accept or decline them on the fly. The durations of the proposed rides are assumed to be i.i.d. with an unknown law, and X_i denotes the duration of the i-th ride. If this ride is accepted, the driver will earn some reward Y_i (that can be either positive or negative, due to costs such as gas, obsolescence of the car, etc.) with expectation $r(X_i)$. We emphasize here that Y_i is not observed before accepting (or actually completing) the i-th ride and that the expected reward function $r(\cdot)$ is unknown to the driver at first. If ride i is accepted, the driver cannot accept any new proposals for the whole duration X_i.

After completing the i-th ride, or after declining it, the driver is on hold, waiting for a new proposal. We assume that idling times – denoted by S_i – are also i.i.d., following some exponential law of parameter λ. An equivalent formulation of the ride arrival process is that proposals follow a Poisson process (with intensity λ) and the driver does not observe ride proposals while occupied.

The driver's objective is to maximize the expected sum of rewards obtained by choosing an appropriate acceptance policy. Given the decisions $(a_i)_{i\geq 1}$ in $\{0,1\}$ (decline/accept), the total reward accumulated by the driver after the first n proposals is therefore equal to $\sum_{i=1}^{n} Y_i a_i$ and the required amount of time for this is equal to $\mathcal{T}_n \coloneqq \sum_{i=1}^{n} S_i + X_i a_i$. This amount of time is random and strongly depends on the policy. As a consequence, we consider that the driver optimizes the cumulative reward up to time T, so that the number of received ride proposals is random and equal to $\theta \coloneqq \min\{n \in \mathbb{N} \mid \mathcal{T}_n > T\}$. In the following, the optimal policy refers to the policy maximizing the expected cumulated reward at time T. The regret is then defined as the difference in expected cumulative rewards, up to time T, between the optimal policy and the considered policy.

3 Contributions

This section summarizes our different contributions. Their complete description can be found in the full version [Boursier et al., 2021]. Using continuous-time dynamic programming principles, we construct an oracle quasi-optimal policy in terms of accepted and rejected rides: it translates into a single optimal threshold for the ratio of the reward to the duration. Any ride with a profitability above this threshold is accepted, and the other ones are declined.

As a benchmark, we first assume that the driver knows the reward function $r(\cdot)$, but ignores the distribution of ride durations. The introduced techniques can be generalized, in the following sections, to further incorporate estimations of $r(\cdot)$. In that case, our base algorithm has a regret scaling as $\mathcal{O}(\sqrt{T})$.

The reward function is then not known to the driver anymore and the reward realizations are assumed to be noisy. To get non-trivial estimation rates, regularity of the reward function is assumed, i.e., (L, β)-Hölder. Modifying the basic algorithm to incorporate non-parametric estimation of r yields a regret scaling as $\mathcal{O}(T^{1-\eta}\sqrt{\ln T})$ where $\eta = \beta/(2\beta+1)$.

Subsequently, our objective is to obtain faster rates, achieved under various structural assumptions on the distribution of ride durations or on the reward function. For instance, we consider margin conditions, which were first used in the context of binary classification [see, e.g., Mammen and Tsybakov, 1999, Tsybakov, 2006], and then introduced in the bandit literature [Perchet and Rigollet, 2013, Goldenshluger and Zeevi, 2009, Weed et al., 2016]. A different assumption used to achieve good rates relates to finite support of the distribution [see e.g., Cesa-Bianchi et al., 2019]. Monotonicity of the profitability function is another assumption in this direction, as it ensures that the optimal policy belongs to some specific subclass, as in [Cesa-Bianchi et al., 2015].

First, we assume that the distribution of ride durations has finite support. We obtain an upper bound on the regret of order $\mathcal{O}(\sqrt{KT\ln T})$, where K is the cardinality of the support. Then we investigate the addition of a margin condition on the distribution. Such assumption gives further information on how difficult it is to distinguish the optimal profitability threshold. The rate obtained is $\mathcal{O}(T^{1-\eta(1+\alpha)}\sqrt{\ln T})$, where $\alpha \in [0,1)$ is a parameter of the margin condition. We then assume that the profitability function is monotone. This ensures the existence of an optimal policy in terms of a threshold for the ride durations – and no longer for the profitability function – for which the rate is $\mathcal{O}(\sqrt{T\ln T})$.

Finally, we show that all our regret bounds are minimax up to poly-logarithmic terms in T and we empirically illustrate our different results on toy examples. Fig. 1 below summarizes our regret bounds with all the considered settings.

4 Related work

Much of the literature on ride hailing focuses on the issue of (surge) pricing by the market platform [see e.g. Özkan and Ward, 2020, Garg and Nazerzadeh, 2020, Besbes et al., 2021]. In contrast to these works that consider drivers with perfect knowledge, we here take the point of view of a driver who learns her environment. As a consequence, this bears similarities with online learning and multi-armed bandit Bubeck and Cesa-Bianchi [2012] as data are gathered sequentially. The main difference with multi-armed bandit or resource allocation problems is that the driver's only resource is her time, which has to be spent wisely and, most importantly, saving it actually has some unknown value. The

Setting	Additional assumption	Incurred regret
known r	-	$\sqrt{T}$
no prior knowledge	r is (L, β)-Hölder	$T^{\frac{\beta+1}{2\beta+1}}\sqrt{\ln(T)}$
finite support of ride duration	$K =$ size of support	$\sqrt{KT \ln T}$
margin condition	r is (L, β)-Hölder margin parameter $\alpha \in [0, 1)$	$T^{1-\frac{\beta(1+\alpha)}{2\beta+1}}\sqrt{\ln T}$
$x \mapsto \frac{r(x)}{x}$ is monotone	-	$\sqrt{T \ln T}$

Figure 1: Order of incurred regrets for different settings.

problem of time allocation actually goes way back.

The driver problem is strongly related to contextual multi-armed bandits, where each arm produces a noisy reward that depends on an observable context. Indeed, the driver faces a one arm contextual bandit problem [Sarkar, 1991], with the crucial difference that the normalized reward associated to a context depends on the whole distribution of contexts (and not just to the current context).

The driver problem is also related to bandits with knapsacks [see Slivkins, 2019, Chapter 10] and even more specifically to contextual bandits with knapsacks, where pulling an arm consumes a limited resource. Time is the resource of the driver here, while both time and resource are well separated quantities in knapsack bandits. Especially, knapsack bandits assume the existence of a null arm, which does not consume any resource. This ensures the feasibility of the linear program giving the optimal fixed strategy. In the driver problem, the null arm (declining the ride) still consumes the waiting time before receiving the next ride proposal. Knapsack bandits strategies can thus not be adapted to the driver problem, which remains solvable thanks to a particular problem structure.

References

Omar Besbes, Francisco Castro, and Ilan Lobel. Surge pricing and its spatial supply response. *Management Sci.*, forthcoming, 2021.

Etienne Boursier, Tristan Garrec, Vianney Perchet, and Marco Scarsini. Making the most of your day: online learning for optimal allocation of time. *arXiv preprint arXiv:2102.08087*, 2021.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Trans. Inform. Theory*, 61(1):549–564, 2015.

Nicolò Cesa-Bianchi, Tommaso Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In *Algorithmic Learning Theory*, pages 247–273. , 2019.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1587–1595, New York, New York, USA, 20–22 Jun 2016. PMLR.

Nikhil Garg and Hamid Nazerzadeh. Driver surge pricing. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, page 501, New York, NY, USA, 2020. Association for Computing Machinery.

Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.*, 19(4):1603–1633, 2009.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

Erhun Özkan and Amy R. Ward. Dynamic matching for real-time ride sharing. *Stoch. Syst.*, 10(1):29–70, 2020.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41(2):693–721, 2013.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *Ann. Statist.*, 19(4): 1978–2002, 1991.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Alexandre Tsybakov. Statistique appliquée, 2006. Lecture Notes.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583, 2016.

Single-index Extreme-PLS regression

Meryem Bousebata[1,2], Geoffroy Enjolras[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.*
[2] *Univ. Grenoble Alpes, CERAG, 38000 Grenoble, France.*
meryem.bousebata@inria.fr, geoffroy.enjolras@grenoble-iae.fr, stephane.girard@inria.fr

Résumé. L'objectif de cette communication est de proposer une nouvelle approche, appelée Single-index Extreme-PLS, pour la réduction de dimension en régression qui soit adaptée aux queues de distributions. Nous nous intéressons à la combinaison linéaire des prédicteurs qui explique au mieux les valeurs extrêmes de la variable réponse dans un contexte de régression inverse non linéaire. La normalité asymptotique de l'estimateur Single-index Extreme-PLS est établie sous des hypothèses modérées. Les performances de la méthode sont évaluées par simulations numériques. Une analyse statistique de données de revenu agricole français, considérant des rendements céréaliers extrêmes, est fournie à titre d'illustration.

Mots-clés. Valeurs extrêmes, Réduction de dimension, Régression inverse non linéaire, Partial Least Squares.

Abstract. The goal of this communication is to propose a new approach, called Single-index Extreme-PLS, for dimension reduction in regression and adapted to distribution tails. The objective is to find a linear combination of predictors that best explain the extreme values of the response variable in a non-linear inverse regression model. The asymptotic normality of the Single-index Extreme-PLS estimator is established under mild assumptions. The performance of the method is assessed on simulated data. A statistical analysis of French farm income data, considering extreme cereal yields, is provided as an illustration.

Keywords. Extreme value, Dimension reduction, Non-linear inverse regression, Partial Least Squares.

1 Introduction

Context. Regression analysis is widely used to study the relationship between a response variable Y and an explanatory p-dimensional vector X starting from an-sample. When p grows, a dimension reduction becomes necessary to show only the most relevant directions of high-dimensional data. There exist a number of statistical models for dimension reduction in regression problems. One of the most popular is Partial Least Squares (PLS) regression, introduced by (Wold, 1975), that combines the characteristics of Principal Component Analysis (PCA) and multiple regression. Its purpose is to find linear combinations of the X coordinates highly correlated with Y. Sliced Inverse Regression

(SIR) is an alternative method for dimension reduction in regression which explores the simplicity of the inverse regression view of X against Y (Li, 1991). It aims at replacing X by its projection onto a subspace of smaller dimension without loss of information. At the same time, there is a growing interest for the modelling of conditional extremes, *i.e.* extremes depending on a covariate. One can mention for instance the estimation of conditional extreme quantiles or more generally, the tail of conditional distributions (Gardes & Girard, 2010). In this communication, we aim to deal with these two lines of works (dimension reduction in regression and conditional extremes) by looking for a linear combination $\beta^t X$ of the covariates that best explains the extreme values of Y. More precisely, we propose a single-index approach to find a direction $\hat{\beta}$ maximizing the covariance between $\beta^t X$ and Y given Y exceeds a high threshold y. This adaptation of the PLS estimator to the extreme-value framework, referred to as Single-index extreme-PLS (SIEPLS), is achieved in the context of a non-linear inverse regression model. In practice, $\hat{\beta}$ allows to quantify the effect of the covariates on the extreme values of Y in a simple and interpretable way. Plotting Y against the projection $\hat{\beta}^t X$ also provides a visual interpretation of conditional extremes. Moreover, working on the pair $(\hat{\beta}^t X, Y)$ should yield improved results for most estimators dealing with conditional extreme values thanks to the dimension reduction achieved thanks to the projection step. From the theoretical point of view, the asymptotic normality of $\hat{\beta}$ is established without linearity or independence requirements.

An inverse model. Let us first consider the following single-index non linear inverse regression model:

(M) $X = g(Y)\beta + \varepsilon$, where X is a p-dimensional random vector, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is the link function and ε is p-dimensional random vector of error. The parameter $\beta \in \mathbb{R}^p$ is an unknown unit vector, ε may depend on Y and g is an unknown function.

Similar inverse regression models were used to establish the theoretical properties of SIR (Bernard-Michel, Gardes & Girard, 2008). Under model (**M**), we aim at estimating β by maximizing the covariance between $\beta^t X$ and Y conditionally on large values of Y. Indeed, roughly speaking, when Y is large, provided the distribution tail of ε is negligible, one has $X \simeq g(Y)\beta$ leading to the approximate single-index model $Y \simeq g^{-1}(\beta^t X)$. Note that the considered model does not require a linear conditional mean or a conditional independence assumption. The paper is organized as follows. In Section 2, the SIEPLS approach is introduced in the framework of a single-index model and heavy-tailed distributions. Some preliminary properties are stated in order to justify the above heuristics from a theoretical point of view. The associated estimator is exhibited in Section 3 and its asymptotic distribution is established under mild assumptions. In Section 4, the performances of the method are investigated through a simulation study and is applied to assess the influence of various parameters on cereal yields collected on French farms.

2 SIEPLS approach

Let us denote by $w(y)$ the unit vector maximizing the covariance between w^tX and Y given that Y exceeds a large threshold y:

$$w(y) = \arg\max_{\|w\|=1} \operatorname{cov}(w^tX, Y|Y \geq y). \tag{1}$$

This optimization problem benefits from a closed-form solution given in the next proposition and obtained by solving the constrained optimization problem using Lagrange multipliers. For all $y \in \mathbb{R}$, let us denote by $\bar{F}(y) = \mathbb{P}(Y \geq y)$ the survival function of Y and the tail-moments, whenever they exist, $m_Y(y) = \mathbb{E}(Y\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}$, $m_X(y) = \mathbb{E}(X\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$, $m_{XY}(y) = \mathbb{E}(XY\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$.

Proposition 1. *Suppose that* $\mathbb{E}\|X\| < \infty$, $\mathbb{E}|Y| < \infty$ *and* $\mathbb{E}\|XY\| < \infty$. *Then, the solution of the optimization problem (1) is:*

$$w(y) = v(y)/\|v(y)\| \text{ where } v(y) = \bar{F}(y)m_{XY}(y) - m_X(y)m_Y(y). \tag{2}$$

Let us note that the solution (2) is invariant with respect to the scaling and location of X. Besides, when ε is centered and independent of Y, we recover the classical PLS framework and it is easily shown that $w(y) = \pm\beta$ for all $y \in \mathbb{R}$. In the following, no assumption is made on the (in)dependence between Y and ε, but additional assumptions on the link function g and the distribution tail of Y are considered:

($\mathbf{A}_1$) Y is a random variable with density function f regularly varying at infinity with index $-\frac{1}{\gamma} - 1$, $\gamma \in (0, 1)$ *i.e.* for all $t > 0$,

$$\lim_{y\to\infty} \frac{f(ty)}{f(y)} = t^{-\frac{1}{\gamma}-1}.$$

This property is denoted for short by $f \in RV_{-1/\gamma-1}$.

($\mathbf{A}_2$) $g \in RV_c$ with $c > 0$.

($\mathbf{A}_3$) There exists $q > 1/(\gamma c)$ such that $\mathbb{E}(\|\varepsilon\|^q) < \infty$.

Let us note that ($\mathbf{A}_1$) implies that $\bar{F} \in RV_{-1/\gamma}$ which is equivalent to assuming that the distribution of Y is in the Fréchet maximum domain of attraction, with extreme-value index $\gamma > 0$, see de Haan & Ferreira (2007). In other words, ($\mathbf{A}_1$) entails that Y has a right heavy-tail. The restriction to $\gamma < 1$ ensures that $\mathbb{E}|Y|$ exists. Assumption ($\mathbf{A}_2$) means that the link function asymptotically behaves like a power function. Finally, ($\mathbf{A}_3$) is a technical assumption which is satisfied for instance by Gaussian distributions.

In order to assess the convergence of $w(y)$ to β as $y \to \infty$, the squared cosine of the angle between the above unit vectors is defined as: $\cos^2(w(y), \beta) = (w(y)^t\beta)^2$. A value close to 0 implies a weak proximity ($w(y)$ is almost orthogonal to β) while a value close to 1 means a high colinearity.

Proposition 2. *Assume* (**M**), ($\mathbf{A}_1$), ($\mathbf{A}_2$) *and* ($\mathbf{A}_3$) *hold with* $\gamma(c+1) < 1$. *Then,*

$$\cos^2(w(y), \beta) = 1 - O\left\{ \left(\frac{1}{g(y)\bar{F}^{1/q}(y)} \right)^2 \right\} \to 0,$$

as $y \to \infty$.

In view of assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$), the function $y \mapsto g(y)\bar{F}^{1/q}(y)$ is regularly varying with index $c - 1/(q\gamma) > 0$ from ($\mathbf{A}_3$). Unsurprisingly, the above convergence rates are large when c is large (*i.e.* the link function is rapidly increasing), q is large (*i.e.* the noise ε is small) or/and γ is large (*i.e.* the tail of Y is heavy). The estimation of $w(y)$ from data distributed from model (**M**) is addressed in the following section.

3 SIEPLS: Population version

Let (X_i, Y_i), $1 \leq i \leq n$ be independent and identically distributed random variables from model (**M**) and let $y_n \to \infty$ as the sample size n tends to infinity. The solution (2) is estimated by its empirical counterpart introducing

$$\hat{v}(y_n) = \hat{\bar{F}}(y_n)\hat{m}_{XY}(y_n) - \hat{m}_X(y_n)\hat{m}_Y(y_n),$$

with $\hat{\bar{F}}$ the empirical survival function and

$$\hat{m}_{XY}(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_Y(y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_X(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y_n\}}.$$

Our main result is the following:

Theorem 1. *Assume* (**M**), ($\mathbf{A}_1$), ($\mathbf{A}_2$) *and* ($\mathbf{A}_3$) *hold with* $2\gamma(c+1) < 1$. *Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$ *and* $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ *as* $n \to \infty$. *Then,*

$$\sqrt{n\bar{F}(y_n)} \left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - \beta \right) \xrightarrow{d} \xi\beta,$$

with $\xi \sim \mathcal{N}(0, \lambda(c,\gamma))$ *and where* $\lambda(c,\gamma)$ *is a constant.*

Assumption $n\bar{F}(y_n) \to \infty$ ensures that the variance of the estimator tends to zero while condition $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ entails that the bias (bounded above by $1/(g(y_n)\bar{F}^{1/q}(y_n))$, see Proposition 2) is asymptotically small compared to the standard deviation $1/\sqrt{n\bar{F}(y_n)}$. Finally, Theorem 1 shows that the estimated direction $\hat{v}(y_n)$ is asymptotically aligned with the true direction β.

4 Numerical results

4.1 Simulated data

We consider a sample of size $n = 1000$ and dimension $p \in \{3, 30\}$ from model (M) with a link function $g(t) = t^c$, $t > 0$, $c \in \{1/4, 1/2, 1, 3/2, 2\}$. The results (available in Bousebata, Enjolras & Girard (2021)) are not reported here for lack of space reasons, they will provided during the presentation.

4.2 Real data

Our approach is applied to data extracted from the Farm Accountancy Data Network (FADN), an annual database of commercial-sized farm holdings. This dataset of $n = 949$ observations contains significant accounting and financial information about French professional farm incomes in 2014. Our goal is to investigate the impact of various factors on farm yields (expressed in quintals per hectare). The response variable Y is the inverse of the wheat yield (in quintals/hectare), as we are interested in the analysis of low yields, and the covariate X includes 12 continuous variables: selling prices (euro/quintal), pesticides, fertilizers, crop insurance purchased, insurance claims, farm subsidies, seeds and seedlings costs, works and services purchase for crops, other insurance premiums, farm income taxes, farmer's personal social security cost (euro/hectare) and temperature average (degree Celsius). A number of visual checks of whether the heavy-tailed assumption makes sense for these data have been implemented (Hill plot and quantile-quantile plot). The estimator SIEPLS $\hat{v}(y_n)$ is computed for each $y_n = Y_{n-k+1,n}$. For the sake of interpretation, we define the conditional correlation between the projected covariate $\hat{v}(y_n)^t X$ and each coordinate $X^{(j)}$ of the covariate as:

$$\rho(X^t\hat{v}(y_n), X^{(j)}|Y \geq y_n) = \frac{\mathrm{cov}(X^t\hat{v}(y_n), X^{(j)}|Y \geq y_n)}{\sigma(X^t\hat{v}(y_n)|Y \geq y_n)\sigma(X^{(j)}|Y \geq y_n)}.$$

Results are depicted on Figure 1 for the 12 considered covariates. Note that the 150 largest inverse wheat yields are mainly consequences of operational costs (fertilisers, pesticides, seeds and seedlings), structural costs (claims, purchase of an insurance policy, farm subsidies, social security cost) and supplementary costs (works and services purchase). This result may be explained by the fact that, in 2014, yields were strongly impacted by production costs, despite mild winter temperatures. Finally, two estimators (linear and non-linear) of the conditional mean $\mathbb{E}(\hat{v}(y_n)^t X|Y)$ have been computed. A positive trend appears for large values of Y in accordance to the inverse regression model (**M**).

Acknowledgements. This work is supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02).

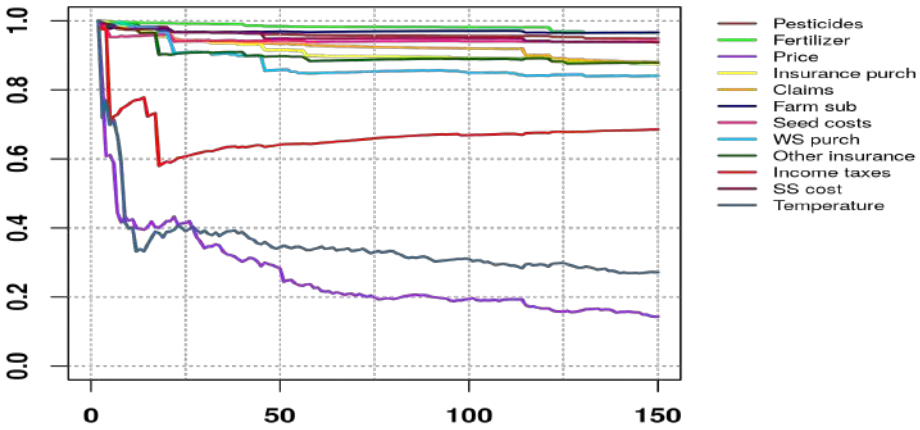


Figure 1: Graph of the estimated conditional correlation function $y \mapsto \rho(X^t\hat{v}(y), X^{(j)}|Y \geq y)$ for $j = 1, \ldots, 12$ (horizontally: number of exceedances k, vertically: conditional correlation estimated by its empirical counterpart using the threshold $y = Y_{n-k+1,n}$).

References

Bernard-Michel, C., Gardes, L., & Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64(3), 982–984.

Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). Regular variation (Vol. 27). *Cambridge university press*.

Bousebata, M., Enjolras, G., & Girard, S. (2021). Extreme Partial Least-Squares regression, *Submitted*, `https://hal.inria.fr/hal-03165399`

de Haan, L., & Ferreira, A. (2007). Extreme value theory: An introduction. *Springer Science & Business Media*.

Gardes, L., & Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2), 177–204.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316–327.

Nelsen, R. B. (2007). An introduction to copulas. *Springer Science & Business Media*.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, 12(S1), 117–142.

The Stochastic block model meets the embedded topic model

Rémi Boutin [1] & Pierre Latouche [1] & Charles Bouveyron [2]

[1] *Université de Paris, MAP 5, UMR 8145, Paris, France*
remi.boutin@u-paris.fr *pierre.latouche@maths.cnrs.fr*
[2] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team*
charles.bouveyron@univ-cotedazur.fr

Résumé. Dans le cadre d'un graphe dont les connexions signifient l'échange de documents textes, nous proposons le ETSBM (Embedded Topics for the Stochastic Block Model) pour détecter des communautés parmi les nœuds en utilisant les thèmes évoqués dans les documents. Sur la base du modèle STBM (Stochastic Topic Block Model), nous remplaçons la détection des topics, reposant sur le modèle LDA (Latent Dirichlet Allocation) par le modèle ETM (Embedded Topic Model) pour bénéficier d'une distribution variationnelle plus souple, et afin d'utiliser des représentations vectorielles du vocabulaire pré-entraînées.

Mots-clés. embeddings, Topic Models, détection de communautés, modèles génératifs, STBM, ETM, LDA, SBM.

Abstract. Considering a graph for which two nodes are linked if and only if they share textual data, we introduce ETSBM (Embedded Topics for the Stochastic Block Model) in order to cluster the nodes using both the links and the topics of the textual data. Based on STBM (Stochastic Topic Block Model), we replaced the topic model block, based on LDA (Latent Dirichlet Allocation), using only count data, with ETM (Embedded Topic Model) to benefit both from the flexible variational distribution and the possibility to use pre-trained embeddings.

Keywords. embeddings, Topic Models, graph clustering, generative models, STBM, ETM, LDA, SBM.

1 Introduction and notations

The STBM model proposed in [C. Bouveyron et al., (2016)] aims at jointly analysing textual data and graph connections in order to perform simultaneously node clustering and topic modelling tasks for each detected community. The analysis of the textual data relies on word counts through LDA (Latent Dirichlet Allocation), see [D. M. Blei et al. (2003)] for more details. In LDA, the word "topic" refers to a distribution over the vocabulary. To overcome some of the limitations of LDA, e.g the bag-of-words representation of documents, ETM (Embedded Topic Model) has been introduced in [A. B. Dieng et al., (2019)].

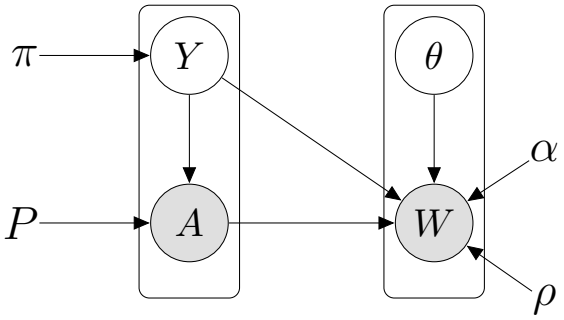


Figure 1: Graphical representation of the Embedded Topics for the Stochastic Block Model. The link between θ and δ is deterministic. In this figure, we give the dependency through θ.

ETM makes use of a deep generative network that jointly learns word and topic representations in a vector space. It can also be used with pre-trained word embeddings, i.e vector representations of words, to incorporate semantic meaning of the words to improve the topic quality. We propose to use ETM instead of LDA within STBM to benefit from those advantages.

In the rest of this paper, we consider a graph with M nodes, for which $A \in \mathbb{R}^{M \times M}$ is the binary adjacency matrix of the graph such that there are no self loop, i.e $A_{ii} = 0$ for any i. Moreover, for any $i, j \in \{1, \dots, M\}$ with $i \neq j$, i is connected to j if and only if A_{ij} equals one. If so, they share a set of documents $W_{ij} := \{W_{ij}^1, \dots, W_{ij}^{D_{ij}}\}$ where D_{ij} denotes the number of documents shared by i and j such that for any $d = 1, \dots, D_{ij}$, W_{ij}^d is a collection of words with size N_{ij}^d, i.e $W_{ij}^d = \{w_{ij}^{d1}, \dots, w_{ij}^{dN_{ij}^d}\}$. V denotes the vocabulary size and the k-th word of the vocabulary will either be represented with a one hot encoded vector or directly with the vocabulary index of the word $w = k$. The embeddings are denoted $\rho \in \mathbb{R}^{d \times V}$ with d the dimension of the vector space and α_k refers to the vector representation of the topic k in the same space for all $k \in \{1, \dots, K\}$, with K the number of topics .

2 Generative model

This section presents our assumptions about the generation of the real data and are represented in Figure 1. The cluster memberships of the nodes are assumed to be independent and identically distributed, following a multinomial distribution of parameter $\pi \in \mathbb{R}_+^Q$ with $\sum_{q=1}^Q \pi_q = 1$ and Q denoting the number of clusters.

$$p(Y \mid \pi) = \prod_{i=1}^{M} \prod_{q=1}^{Q} \pi_q^{Y_{iq}}. \tag{1}$$

Given each node's cluster, the connections between the nodes are assumed to be independent:

$$p(A \mid P, Y) = \prod_{i \neq j}^{M} \prod_{q,r}^{Q} P_{qr}^{Y_{iq} Y_{jr} A_{ij}}. \tag{2}$$

For any pair of clusters (q, r), the prior on the topic proportions $\theta_{q,r}$ is a logistic normal distribution:

$$\delta_{qr} \sim \mathcal{N}(0_K, \mathrm{I}_K), \qquad \theta_{qr} = \mathrm{softmax}\left(\delta_{qr}\right).$$

with $\mathrm{softmax}(\delta) = \left(\frac{e^{\delta_1}}{\sum_k^K e^{\delta_k}}, \ldots, \frac{e^{\delta_K}}{\sum_k^K e^{\delta_k}} \right)$ for any $\delta \in \mathbb{R}^K$.

If two nodes i, j are connected and if they are in the clusters q and r respectively, the n-th word of the d-th document is assumed to be distributed according to a mixture of topics:

$$p(w_{ij}^{dn} \mid \theta_{qr}, Y_{iq} Y_{jr} A_{ij} = 1) = \sum_{k=1}^{K} \theta_{qrk} \, \mathrm{softmax}(\rho^\top \alpha_k)_{|w_{ij}^{dn}}. \tag{3}$$

Finally, given each node's cluster membership Y_i, marginalizing the topic proportions gives the following:

$$\begin{aligned} p(W \mid Y, A, \alpha, \rho) &= \prod_{i \neq j}^{M} \left\{ \prod_{q,r}^{Q} p\left(w_{ij} \mid Y_{iq} Y_{jr} A_{ij} = 1, \alpha, \rho\right)^{Y_{iq} Y_{jr}} \right\}^{A_{ij}} \\ &= \prod_{i \neq j}^{M} \left\{ \prod_{q,r}^{Q} \left(\int p\left(\delta_{qr}\right) \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^d} p\left(w_{ij}^{dn} \mid \delta_{qr}, Y_{iq} Y_{jr} A_{ij} = 1\right) d\delta_{qr} \right)^{Y_{iq} Y_{jr}} \right\}^{A_{ij}}. \end{aligned} \tag{4}$$

3 Inference

We aim at maximizing the **joint log-likelihood** $\log p(A, Y, W \mid P, \pi, \alpha, \rho)$ of our model with respect to the parameters P, π, α, ρ and to the latent variable Y given by:

$$\log p(A, Y, W \mid P, \pi, \alpha, \rho) = \log p(A \mid Y, P) + \log p(W \mid A, Y, \rho, \alpha) + \log p(Y \mid \pi). \tag{5}$$

While optimizing with respect to P and π is a direct application of the SBM model, the *"topic term"* $\log p(W \mid A, Y, \rho, \alpha)$ is more challenging. We use a variational-inference approach to infer the parameters which allows to split the log-likelihood between the

ELBO (expected lower bound) and the Kullback-Leibler divergence between the posterior and a variational distribution $R_\nu(\delta; w)$:

$$\log p(W \mid A, Y, \rho, \alpha) = \underbrace{\mathbb{E}_{R_\nu}\left[\log \frac{p(W, \delta \mid A, Y, \rho, \alpha)}{R_\nu(\delta; w)}\right]}_{=:\mathscr{L}(\alpha,\rho,Y;\nu)\ \text{(ELBO)}} + \mathrm{KL}\Big(R_\nu(\delta; w) \mid\mid p(\delta \mid w)\Big). \tag{6}$$

We also make use of amortized inference as in [D. P. Kingma and M. Welling (2014)] and [D. J. Rezende et al. (2014)] which allows the number of variational parameters not to grow linearly with the number of observations. This differs from the usual mean-field assumption since it requires to build a mapping that efficiently encodes the data into a vector space of dimension d, and using a single variational parameter to parametrize this mapping, e.g a neural network. Thus, the following variational distribution is used:

$$R_\nu(\delta; W) = \prod_{qr}^{Q} \mathcal{N}(\delta_{qr}; \mu_{qr}^\nu, \sigma_{qr}^\nu), \tag{7}$$

with $\mu_{qr}^\nu = f_1(W_{qr}; \nu)$ and $\sigma_{qr}^\nu = f_2(W_{qr}; \nu)$, where $f_1(\cdot, \nu)$ and $f_2(\cdot, \nu)$ are the model encoders sharing the hidden layer parametrized by ν, $W_{qr} := \{W_{ij} : Y_{iq}Y_{jr}A_{ij} = 1\}$. The ELBO can once again be split in two terms allowing easier computations:

$$\begin{aligned}
\mathscr{L}(\alpha, \rho, Y; \nu) &= \int_\delta R_\nu(\delta; W) \log\Big\{p\left(W \mid Y, A, \delta, \alpha, \rho\right) \frac{p(\delta)}{R_\nu\left(\delta; W\right)}\Big\} d\delta && (8)\\
&= \mathbb{E}_{R_\nu}\left[\log p\left(W \mid Y, A, \delta, \alpha, \rho\right)\right] - \mathrm{KL}\Big(R_\nu(\delta; W) \mid\mid p(\delta)\Big). && (9)
\end{aligned}$$

The second term is the Kullback-Leibler divergence between two Gaussians and has a closed form. Let's rewrite the first term to exhibit an estimator of the integral:

$$\mathbb{E}_{R_\nu}\left[\log p\left(W \mid Y, A, \delta, \alpha, \rho\right)\right] = \sum_{i \neq j}^{M} \sum_{q,r}^{Q} A_{ij} Y_{iq} Y_{jr} \mathbb{E}_{R_\nu}\Big[\log \underbrace{p\left(w_{ij} \mid \delta_{qr}, \alpha, \rho\right)}_{T_{ij}^{\delta_{qr}}}\Big], \tag{10}$$

with

$$\log T_{ij}^{\delta_{qr}} = \sum_{d=1}^{D_{ij}} \sum_{n=1}^{N_{id}^d} \log\left\{\sum_{k=1}^{K} \mathrm{softmax}(\delta_{qr})_{|k}\, \mathrm{softmax}(\rho^\top \alpha_k)_{|w_{ij}^{dn}}\right\}. \tag{11}$$

In order to get an estimator of the intractable quantity $\mathbb{E}_{R_\nu}\left[\log T_{ij}^{qr}\right]$, we also make use of a Monte-Carlo approximation. For all pairs (q, r), we draw S samples such that:

$$\epsilon^s \sim \mathcal{N}(0, I_K), \qquad \delta_{qr}^s = \mu_{qr}^\nu + \sigma_{qr}^\nu \odot \epsilon^s, \tag{12}$$

where $\odot$ is the element-wise product, i.e $\delta^s_{qrk} \sim \mathcal{N}\left(\mu^\nu_{qrk}, (\sigma^\nu_{qrk})^2\right)$. Thus, for all pairs (i, j) and (q, r), the following estimator holds:

$$C^{qr}_{ij} = S^{-1} \sum_{s=1}^{S} \log T^{\delta^s_{qr}}_{ij}. \tag{13}$$

4 Optimization

A unified version of the optimisation can be found in algorithm 1 based on a classification VEM procedure, see [C. Bouveyron et al., (2016)] for more details.

Optimization of π and P: the Graph-step Following the original paper, maximising (5) with respect to π and P comes down to maximizing $\log p(Y \mid \pi)$ and $\log\ p(A \mid Y, P)$ respectively. Adding the constraint $\sum_{q=1}^{Q} \pi_q = 1$ and setting the Lagrangian derivatives to zero gives the following equations:

$$\hat{\pi}_q = \frac{1}{M} \sum_{i=1}^{M} Y_{iq}, \qquad \hat{P}_{qr} = \frac{\sum_{i \neq j} A_{ij} Y_{iq} Y_{jr}}{\sum_{i \neq j} Y_{iq} Y_{jr}}. \tag{14}$$

Optimization of ρ and α: the NLP-step Considering that π, P and Y are held fixed, we now focus on maximizing the ELBO with respect to ρ and α. If the embeddings are not pre-trained, both ρ and α are learned. Otherwise, ρ stays fixed and only α is learned. The optimization uses the Variational Bayes approach described in [D. P. Kingma and M. Welling (2014)]. It relies on a gradient-based optimization of the ELBO $\mathscr{L}(\alpha, \rho, Y; \nu)$ using the re-parametrization trick and the amortized inference. For the optimization to be scalable, a mini-batch optimization is performed.

Optimization of Y: the Clustering-step Finally, given the other parameters, a greedy optimization is performed on Y. We use the same approach as in [C. Bouveyron et al., (2016)]. Considering all the Y_j fixed for all $j \neq i$, we cycle through the possibilities of clusters for Y_i and assign it with the one maximizing the ELBO. This procedure is repeated for all nodes i, using the already updated clusters of the previous nodes.

Algorithm 1: Optimization algorithm of the ETSBM

Initialize $\rho, \alpha, \theta, \nu$ randomly and Y_i with a K-means on matrix A ;
Initialize P and π using (14);
For all pair (q, r), draw S samples $\delta_{qr}^s \sim q(\delta_{qr} \mid \nu, w_{qr})$;
For all i, j, (q, r) set C_{ij}^{qr} using (13) and $L = \sum_{i,j;j\neq i}^{M} \sum_{q,r}^{Q} A_{ij} Y_{iq} Y_{jr} C_{ij}^{qr}$;
while $\hat{Y} \neq Y$ **do**
  Perform the Graph-step then the NLP-step;
  Clustering step
    For each pair (q, r) **draw** S **samples** $\delta_{qr}^s \sim q(\delta_{qr} \mid \nu, \tilde{w}_{qr})$;
    for $i = 1, \ldots, M$ **do**
      for $q = 1, \ldots, Q$ **do**
        Set $\tilde{Y}_{iq} = 1$;
        Compute $\tilde{C}_{ij}^{qr}$ then $\tilde{L}$;
        if $\tilde{L} > L$ **then** $Y_i = \tilde{Y}_i$ and $L = \tilde{L}$;
      end
    end
  end
end

Further works This section is under current investigation. We will conduct a set of simulations based on the ones in [C. Bouveyron et al., (2016)]. An application will be carried out on a dataset of academic papers and a co-authors network to detect both groups of authors and underlying topics.

References

[C. Bouveyron et al., (2016)] C. Bouveyron, P. Latouche and R. Zreik (2016), The stochastic topic block model for the clustering of vertices in networks with textual edges, *Statistics and Computing*, 28, 11–31.

[D. P. Kingma and M. Welling (2014)] D. P. Kingma and M. Welling (2014), Auto-Encoding Variational Bayes, *ICLR 2014*.

[D. M. Blei et al. (2003)] D. M. Blei, A. Y. Ng and M. I. Jordan (2003), Latent Dirichlet Allocation, *JMLR.org*, 3, 993–1022.

[A. B. Dieng et al., (2019)] A. B. Dieng, F. J. R. Ruiz and D. M. Blei (2019), Topic Modeling in Embedding Spaces, *CoRR 2019*

[D. J. Rezende et al. (2014)] D. J. Rezende, S. Mohamed and D. Wierstra (2014), Stochastic Backpropagation and Approximate Inference in Deep Generative Models, *PMLR 2014*

Modèle partiellement censuré pour l'aide à l'estimation de la prévalence dans le cadre de la pandémie SARS-CoV-2

Vincent Brault[1,4] & Bastien Mallein[2,4] & Jean-François Rupprecht[3,4]

[1] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
vincent.brault@univ-grenoble-alpes.fr
[2] *Université Sorbonne Paris Nord, LAGA, UMR 7539, F-93430, Villetaneuse, France.*
mallein@math.univ-paris13.fr
[3] *Aix Marseille Univ, CNRS, Centre de Physique Théorique, Turing Center for Living Systems, Marseille, France.*
rupprecht@cpt.univ-mrs.fr
[4] *Members of the GROUPOOL & MODCOV19 initiatives.*

Résumé. Afin de limiter la propagation du SARS-CoV-2, l'utilisation du poolage (ou tests groupés) est une solution qui est déjà utilisée dans de nombreux pays (Israël, États-Unis, Royaume Uni...). Le principe étant de mélanger plusieurs échantillons, il a été montré qu'il y a une augmentation du taux de faux négatifs dépendant de la distribution de la charge virale dans la population. Dans cet exposé, nous proposons une modélisation pour estimer cette distribution. Pour ce faire, nous introduisons un modèle de mélange gaussien avec une censure totale ou partielle afin d'estimer au mieux les données.

Mots-clés. Modèle de mélanges, censure, SARS-CoV-2, poolage

Abstract. During the epidemic SARS-CoV-2, pooling is used to aid diagnosis or estimate prevalence. In practice, samples are mixed which dilutes the viral load and may involve a greater number of false negatives. However, the influence of the dilution depends on the distribution of the viral load in the population. In this talk, we propose a model to estimate this distribution. We introduce a Gaussian mixture model with total or partial censorship in order to best estimate the data.

Keywords. Mixture Model, Censorship, SARS-CoV-2, pooling

1 Introduction

Dans le cadre de maladies, notamment contagieuses, il est important de surveiller la prévalence (c'est-à-dire le nombre de personnes infectées) afin de mieux la maîtriser (voir par exemple Salje et al. 2020). Pour la pandémie de SARS-CoV-2, certains malades sont asymptomatiques et les autres peuvent mettre du temps à développer les symptômes (temps durant lequel ils sont contagieux). Une solution pour réussir la détection précoce

est de faire une campagne de tests à grande échelle (voir Lavezzo et al (2020)). Une solution pour pouvoir anticiper des problèmes de pénurie de réactifs ou des engorgements dans les laboratoires est d'utiliser la méthode de poolage dont le principe est de mélanger les échantillons et de tester uniquement ce mélange (voir par exemple Dorfman (1943) pour une application dans le cas de la syphilis). Le principe est que, dans le cadre d'un test parfait, si le résultat est négatif alors aucun patient du groupe n'est contaminé et si le résultat est positif, il y a au moins un malade dans le groupe. Dans le cas du SARS-CoV-2, des recherches ont montré empiriquement que cette méthode accroît les capacités de test en maintenant un haut degré de sensitivité (voir par exemple Ben-Ami (2020)). Dans leur article, Brault et al. (2021) proposent une modélisation afin de comprendre l'influence du poolage dans le taux de faux négatifs des tests *reverse transcription quantitative polymerase chain reaction* (RT-qPCR) ; les plus utilisés pour l'instant pour détecter la charge virale (voir Corman et al. (2020)). Pour ce faire, ils ont besoin de connaître la distribution de la charge virale au sein de la population.

Dans cet exposé, nous présentons la partie de l'article de Brault et al. (2021) portant sur cette estimation. Pour ce faire, nous commencerons par rappeler le principe général utilisé dans les tests RT-qPCR puis nous expliciterons le modèle avec une censure partielle ou totale et la généralisation à un modèle de mélange. Nous terminerons par une application sur des données simulées mimant des données réelles.

2 Contexte

Pour estimer si un individu est contaminé ou pas, le processus consiste à prélever un échantillon (soit dans le nez, soit dans la salive) puis de faire une manipulation pour que la séquence d'ARN cible soit transcrite en ADN. L'échantillon est ensuite placé dans une machine PCR mesurant la concentration de brins d'intérêt en le rendant fluorescent ; un réactif est ajouté pour doubler à peu près le nombre de brins d'intérêt à chaque cycle (le coefficient multiplicateur diminue à partir d'un certain nombre de cycles). Ainsi, s'il y a une charge virale, il est censé y avoir une fluorescence à un moment : plus elle se fait tardivement, plus la charge virale est faible.

En pratique, si la machine fonctionnait éternellement, une fluorescence risquerait d'apparaître à un moment à cause d'une *impureté* qui pourrait être prise pour de la charge virale ce qui impliquerait un faux positif. Pour contrer cela, la machine est programmée pour ne faire qu'un nombre réduit de cycles ; dans ce cas, les faibles charges virales risquent de ne pas être détectées (voir le schéma de la figure 1). Dans le cas de la détection du SARS-CoV-2, il est supposé que les machines sont calibrées de telle sorte qu'il n'y ait aucun faux positif. Ceci implique qu'il y a certainement des faux négatifs.

Dans leur article, Brault et al. (2021) s'intéressent à l'influence de la dilution imposée dans le cadre d'une procédure de poolage sur l'augmentation du nombre de faux négatifs puisque la charge virale initiale va diminuer. La modélisation proposée représentant le

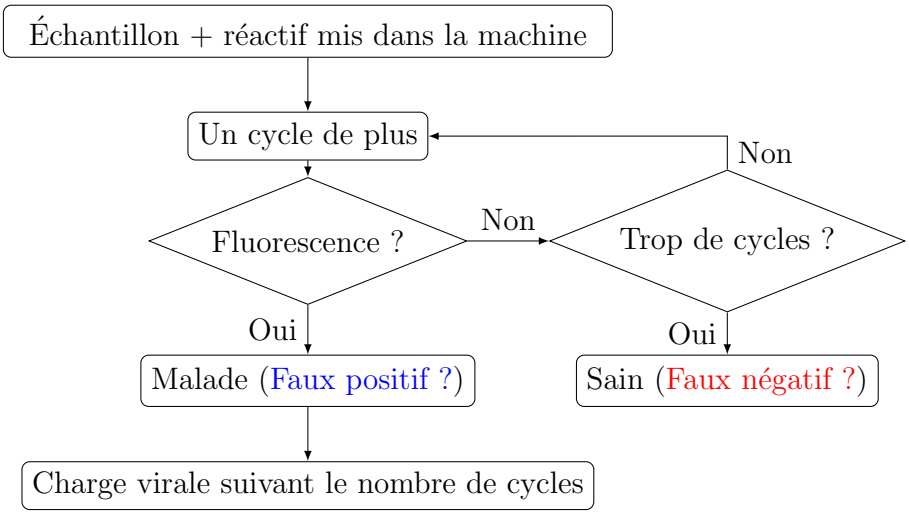


Figure 1: Représentation schématique du processus pour estimer si un individu est contaminé ou pas en fonction de sa charge virale initiale. Un faux positif (déclaration à tord que l'individu est malade) est obtenu si une impureté est prise pour une charge virale (dans le cas où trop de cycles sont effectués). Un faux négatif correspond à un individu supposé sain alors qu'il possède une charge virale trop faible pour être détectée par le nombre limite de cycles proposés.

nombre de cycles en fonction de la charge virale implique une translation de la distribution du nombre de cycles en logarithme du nombre de personnes testées dans un groupe.

Il apparaît donc important d'estimer la distribution du nombre de cycles pour connaître la proportion du nombre de faux négatifs en plus que cela entraîne. Notons que les valeurs des nombres de cycle ne sont pas entières à cause d'une uniformisation des résultats obtenus par différentes machines calibrées légèrement différemment ; ceci peut également avoir une légèrement influence sur le seuil théorique du nombre maximum de cycles.

3 Modèle de mélange avec une censure fixée

Afin d'estimer la distribution du nombre de cycles, nous avons repris l'histogramme obtenu par Jones et al. (2020) sur une population de 3 712 individus. Comme les données ne sont pas publiques, nous avons simulé des échantillons de telle sorte à obtenir les mêmes histogrammes ; pour chaque barre, nous avons pris une loi uniforme entre les deux bornes (voir la gauche de la figure 2).

La forme de l'histogramme laisse penser que nous sommes en présence d'un modèle de mélange. Nous avons donc utilisé un algorithme EM (*Expectation Maximisation* ; voir Dempster et al. (1977)) et un critère BIC (*Bayesian Information Criterion* ; voir Schwarz

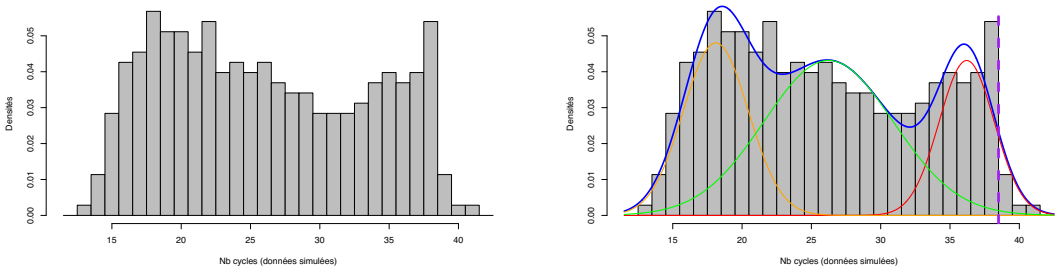


Figure 2: Histogramme des données simulées mimant l'histogramme de Jones et al. (2020). À droite, nous avons représenté l'estimation par un modèle de mélange avec trois composantes. Un trait vertical en pointillé a été ajouté pour signifier l'emplacement probable de la censure.

(1978)) pour estimer le nombre de composantes et les paramètres. Comme nous simulons les données, nous avons relancé 100 fois la procédure. Nous obtenons dans 5% des cas deux composantes et dans 95% des cas trois composantes. Sur la droite de la figure 2, nous avons représenté une estimation avec 3 composantes : nous observons que les deux composantes de gauche estiment plutôt bien la distribution mais la composante la plus à droite semble obligée de descendre prématurément. Ceci peut-être dû à la censure qui se situerait aux alentours de 38,5 cycles (voir le trait vertical en pointillé sur la droite de la figure 2).

Nous présentons dans la section suivante le modèle gaussien (partiellement) censuré que nous proposons d'utiliser comme base pour le modèle de mélange de la section 3.2.

3.1 Modèle gaussien (partiellement) censuré

Pour prendre en compte cette censure, notée d_{cens}, nous proposons deux modèles se basant sur un modèle gaussien. Dans les deux cas, nous supposons que la distribution de la charge virale suit une loi gaussienne mais les observations à droite de la censure ne sont pas forcément observées.

Dans le modèle à censure partielle, noté $\mathcal{CN}_{d_{cens}}(\mu, \sigma, q)$, nous supposons que l'observation dépend qu'une probabilité $q \in]0; 1]$. Ainsi, nous obtenons la densité suivante :

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{cens})} \left[1 + (q-1)\mathbb{1}_{\{x > d_{cens}\}}(x)\right].$$

Dans le modèle à censure totale, noté $\mathcal{CN}_{d_{cens}}(\mu, \sigma, 0)$, nous supposons qu'aucune observation n'est faite après la censure. Ainsi, nous obtenons la densité suivante :

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{cens})} \mathbb{1}_{\{x \leq d_{cens}\}}(x).$$

Remarquons que si d_{cens} tend vers l'infini ou q vaut 1, nous retrouvons un modèle gaussien standard. Comme ces lois appartiennent à la famille des lois exponentielles, le modèle est identifiable et l'estimateur du maximum de vraisemblance est asymptotiquement normal. Néanmoins, il n'existe pas de forme analytique pour l'estimateur du maximum de vraisemblance et il est nécessaire d'utiliser des algorithmes d'optimisation.

3.2 Modèle de mélange gaussien (partiellement) censuré

Pour l'estimation dans le cadre d'un modèle de mélange, nous avons dû faire plusieurs hypothèses :

- D'abord, nous avons supposé que la censure était la même pour toutes les composantes. Ce choix est motivé par le fait que l'emplacement de la censure dépend de la machine et pas de la charge virale.

- Ensuite, nous avons supposé que la probabilité q dépend de la composante (comme pour la moyenne et la variance) pour le moment. Il serait intéressant d'étudier le cas où la probabilité q est commune à toutes les composantes, ce qui semblerait plus proche de la réalité, mais la maximisation est alors plus compliquée.

Sur la figure 3, nous avons représenté les résultats obtenus pour les deux modélisations (en lignes) en prenant deux ruptures.

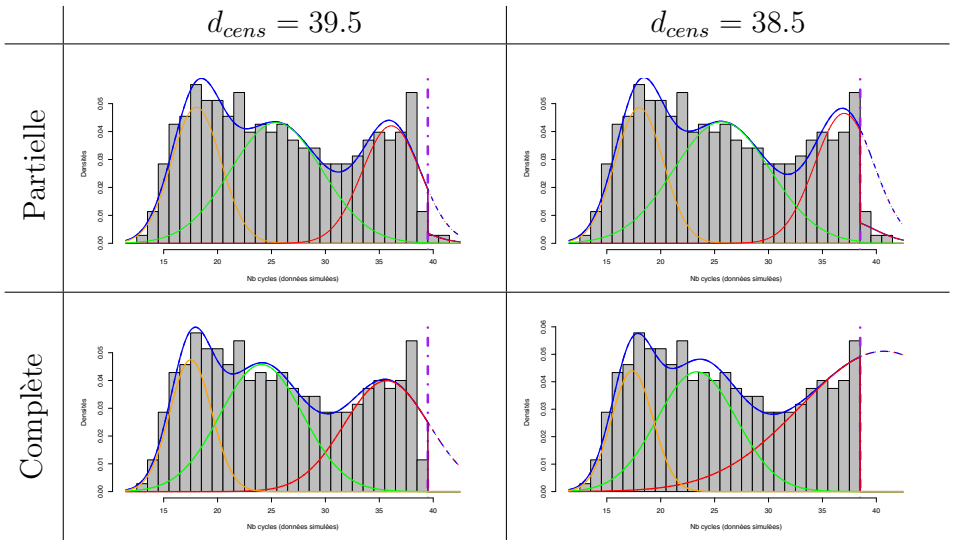


Figure 3: Représentation des densités obtenues en fonction du modèle choisi (en ligne) et de la position de la censure (en colonne).

Nous observons que les deux composantes de gauche ont globalement les mêmes estimations quelque soit le modèle. La densité de droite (plus proche de la rupture) va avoir un moyenne et une variance qui vont se décaler vers la droite lorsque nous rapprochons la rupture du pic de 38.5 avec une plus grosse variance pour le modèle totalement censuré.

4 Conclusion

Dans cet exposé, nous présenterons les différents résultats obtenus sur les simulations et sur les données réelles.

Bibliographie

Ben-Ami R., Klochendler A., Seidel M., Sido T., Gurel-Gurevich O., Yassour M., Meshorer E., Benedek G., Fogel I., Oiknine-Djian E., et al. *Large-scale implementation of pooled rna extraction and rt-pcr for sars-cov-2 detection.* Clinical Microbiology and Infection, 26(9):1248–1253, 2020.

Brault V., Mallein B. et Rupprecht J.-F., *Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance* accepté dans PLOS Computational Biology, 2021.

Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.* Euro-surveillance. 2020;25(3):1–8. doi:10.2807/1560-7917.ES.2020.25.3.2000045.

A. P. Dempster, N. M. Laird, and D. B. Rubin. *Maximum likelihood from incomplete data via the em algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Dorfman R. *The Detection of Defective Members of Large Populations.* The Annals of Mathematical Statistics. 1943

T. C. Jones, B. Mühlemann, T. Veith, G. Biele, M. Zuchowski, J. Hoffmann, A. Stein, A. Edelmann, V. M. Corman, and C. Drosten. *An analysis of sars-cov-2 viral load by patient age.* medRxiv, 2020.

Lavezzo E, Franchin E, Ciavarella C, Cuomo-dannenburg G, Barzon L, Sciro M, et al. *Suppression of COVID-19 outbreak in the municipality of Vo', Italy.* medRxiv. 2020; p. 1–23. doi:10.1101/2020.04.17.20053157.

Salje, H., Kiem, C. T., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Cauchemez, S. (2020). *Estimating the burden of SARS-CoV-2 in France.* Science, 369(6500), 208-211.

G. Schwarz. *Estimating the dimension of a model.* The annals of statistics, 6(2):461–464, 1978.

Important variables are Game-Changers: Revisiting Shapley Values for explaining Black-Box models

Salim Ibrahim Amoukou [1] & Nicolas Brunel [2]

[1] *Stellantis et Université Paris Saclay, CNRS, Laboratoire de Mathématiques et Modélisation d'Evry, salim.ibrahim-amoukou@universite-paris-saclay.fr*
[2] *Université Paris Saclay, CNRS, ENSIIE Laboratoire de Mathématiques et Modélisation d'Evry, nicolas.brunel@ensiie.fr*

Résumé. L'explicabilité des modèles de Machine Learning est un domaine très actif, car il est un vecteur important de l'acceptabilité des algorithmes d'Intelligence Artificielle. Parmi les techniques récemment proposées, les valeurs de Shapley émergent comme un indicateur de référence, car il fournit une explication additive des prédictions. Cependant les valeurs de Shapley utilisées actuellement peuvent être empreintes d'erreurs d'estimation, et sensibles à la présence de variables peu importantes. Nous avons développé un algorithme qui permet de calculer les "same decision probabilities" qui mesurent la probabilité de garder la même décision en ne fixant qu'une partie des variables prédictives. Ceci nous permet d'introduire un nouveau jeu coopératif qui permet de montrer que les variables qui contribuent le plus à cette stabilité sont des variables importantes du modèle. Nous illustrons les concepts proposés sur un modèle graphique.

Mots-clés. Valeurs de Shapley, Explicabilité, Apprentissage Automatique, Sélection de variables, Importance de Variables.

Abstract. The explainability of Machine Learning models is a very active field, because it is an important vector of the acceptability of AI algorithms. Among the recently proposed techniques, Shapley Values have emerged as a gold-standard, as it provides an additive explanation of predictions. However, Shapley Values are often prone to estimation errors, and are sensitive to the presence of unimportant variables. We have introduced a new computation algorithm which allows us to compute also the "Same Decision Probability", which measures the probability of keeping the same decision by fixing a subset of the predictor variables. Our main contribution is the introduction of a new cooperative game that shows that the variables who acts as game-changer are the important variables of the model. We illustrate our findings on a graphical model.

Keywords. Shapley values, Explainable AI, Variable Importance, Variable Selection, Machine Learning.

1 Introduction

This work addresses the problem of interpretability of Machine Learning models. Despite a growing use of machine learning in applications and real life problems, a significant part of previous academic works was dedicated to the improvement of the prediction capabilities or computational efficiencies of ML Models until the recent years. The objective of the very active and recent field of Explainable AI (XAI) aims at developing tools that could provide better insights in the important variables, at a global or at a local level. While statistical models are often based on some testable assumptions, or might be interpretable by design, there is a need for development of model-agnostic importance measures for ML models, in order to be able to understand the differences between very diverse models and to perform some sort of variables selection. Among the most used local measures, the Shapley Values (SV) comes from cooperative game theory and evaluates the "fair" contribution of a variable $X_i = x_i$ in a prediction [1]. One of the main interest of Shapley values, is that they provide an additive (decomposition) explanation of the prediction, which makes it relatively easy to understand. While Shapley Values are considered as one of the state-of-the-art methodology, several critics have been addressed concerning the computational complexity and the approximations needed, or the difficulty to relate them to other interpretable frame work, such as causality [2, 3]. We recall in the next section the definition of the Shapley Values. In section 3, we introduce a measure of stability, called "same decision probability" that we use for computing the importance of group of variables. Finally, in order to obtain a reliable estimate of the importance of a variable we define a new cooperative game where we can define "Swing Shapley Values" that are different than the standard ones introduced in [1]. Finally, we show that the variables that perform as game-changer when we consider swinging coalitions, are important variables. An example on realistic data illustrates our findings and conclusions.

2 Feature attribution and Shapley Values

We recall in this section the definitions and main properties of Shapley Values.

2.1 Shapley values for explaining Black-Box models

For any group of variables $\boldsymbol{X}_S = (X_i)_{i\in S}$, with any subset $S \subseteq [\![1, p]\!]$, we define the reduced predictors as

$$f_S(\boldsymbol{x}_S) \triangleq E\left[f\left(\boldsymbol{X}\right)|\boldsymbol{X}_S = \boldsymbol{x}_S\right]. \tag{2.1}$$

The SV for local interpretability at $\boldsymbol{x}$ are based on a cooperative game with the value function $v(f; S) \triangleq f_S(\boldsymbol{x}_S)$ (a value function is a function from 2^p set to $\mathbb{R}$). For any coalition of variables $C \subseteq [\![1, p]\!]$ and $k \in [\![1, p - |C|]\!]$, we denote the set $\mathcal{S}_k(C) =$

$\{S \subseteq [\![1,p]\!] \setminus C | \, |S| = k\}$: the SV of the coalition C is defined as

$$\phi_C(f;\boldsymbol{x}) = \frac{1}{p-|C|+1} \sum_{k=0}^{p-|C|} \frac{1}{\binom{p-|C|}{k}} \sum_{S \in \mathcal{S}_k(C)} (f_{S\cup C}(\boldsymbol{x}_{S\cup C}) - f_S(\boldsymbol{x}_S))) \tag{2.2}$$

The definition (2.2) of the SV is a straightforward extension of the standard SV of a single variable (or player) to a group of variables. The standard SV is recovered with $C = \{i\}$ for $i \in [\![1,p]\!]$. A reference code for computing SV is the Python Open source library SHAP[1], that implements various approximation algorithms.
The great benefit of the Shapley Values is the so-called additive explanation:

$$f(\boldsymbol{x}) - E[f(\boldsymbol{X})] = \sum_{i=1}^{p} \phi_i(f,\boldsymbol{x}) \tag{2.3}$$

which permits to measure directly the influence of the variable X_i on the prediction. A common classical criticism is about the effective estimation of the expectations needed in the SV computation that is statistically challenging and combined with an exponential complexity. We focus in that paper on tree-based models as the computational cost can be made polynomial and the statistical problem is easier to address [4].

2.2 Closed-form expressions for reduced predictors

The computation of the SV uses all the conditional expectations $E\left[f(\boldsymbol{X})|\boldsymbol{X}_S = \boldsymbol{x}_S\right], S \subseteq [\![1,p]\!]$. While it is difficult in general, the paper [4] introduce a recursive algorithm that reads sequentially and recursively the different nodes. In practice, the conditional expectations need to be estimated from the training or the test set. With tree-structured models, we can have efficient algorithms for computing in closed-form conditional expectations and SV. We assume that we have a tree with M leafs $L_1, \ldots, L_M$ based on the variables $X_1, \ldots, X_p$ (continuous or qualitative), the predictor f is a tree $f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$. The reduced predictor is $f_S(\boldsymbol{x}_S) = \sum_{m=1}^{M} f_m P_X\left(L_m \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$ showing that the only challenge is the computation of the conditional probabilities. We have implemented in a Python package *Active Coalition of Variables*[2] an efficient algorithm for computing more accurate conditional probabilities and Shapley Values for tree-based models.

3 A new game for Variable Importance

In general, we are not only interested in computing feature importance $\phi_i(f,\boldsymbol{x})$, we also want to identify the group of variables $X_i, i \in S$ that best explains $\boldsymbol{x}$ and the group of

[1] https://github.com/slundberg/shap
[2] https://github.com/salimamoukou/acv00

uninformative variables $\boldsymbol{X}_i, i \in \bar{S}$. Therefore, several papers [5, 6, 7] suggest to use SV as a heuristic for feature selection, but as proved in [2], the magnitude of SV of variables do not necessarily correspond to relevant variables. Indeed, a variable can have a low influence but paradoxically, it can have at the same time a high $\phi_i(f, \boldsymbol{x})$. So we need to filter the noisy variables.

3.1 Same Decision Probability and game changer

Our methodology for identifying the most important features is based on the Same Decision Probability (SDP) criterion, introduced in [8], and that can be computed for tree-based models in the *Active Coalition of Variables* library.

Definition 3.1 (**Same Decision Probability of a classifier**). *Let $f : \mathcal{X} \longrightarrow [0, 1]$ a probabilistic predictor and its classifier $C(\boldsymbol{x}) = \mathbb{1}_{f(\boldsymbol{x}) \geq T}$ with threshold T, the Same Decision Probability of coalition $S \subset [\![1, p]\!]$, w.r.t $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$ is*

$$SDP_S\left(C; \boldsymbol{x}\right) = P\left(C(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) = C(\boldsymbol{x}) \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$$

SDP gives the probability to keep the same decision $C(\boldsymbol{x})$ when we do not observe the variables $\boldsymbol{X}_{\bar{S}}$. The higher is the probability, the better is the explanation based on S. We introduce a new cooperative game that will put emphasis on the game-changers i.e on the variables that make the decision becoming stable when they enter into a coalition.

Definition 3.2 (**Swing Shapley Values**). *Let f a model, $\boldsymbol{x}$ an instance, $SDP_S(f; \boldsymbol{x})$ the same decision probability of coalition S. We define the new cooperative game with value function:*

$$v_{SDP,\pi}(f; S) = \begin{cases} 1, & \text{if } SDP_S(f; \boldsymbol{x}) \geq \pi \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

For this game, we can also compute the corresponding Shapley Value (denoted Swing-SV) in order to compute the overall contribution of a variable to the game induced by the value function (3.1). The Swing-SV $\phi_i^{SDP,\pi}$ of variable X_i is then computed by replacing $f_S(\boldsymbol{x}_S)$ by $v_{SDP,\pi}(S)$ in the standard definition of a Shapley Value (2.2). A coalition with value zero is called a "losing coalition" and with value one a "winning coalition". If a player's entry into a coalition changes the value from losing to winning, then the player's contribution is one, otherwise zero. A coalition S is said to be a *swing* for player i if S is losing but $S \cup i$ is winning. Therefore, a high Swing-SV $\phi_i^{SDP,\pi}$ implies that the variable X_i generates a lot of swings and is a game-changer; i.e this variable permits to retrieve significantly the original prediction. However, it should be noted that the SV $\phi_i^{SDP,\pi}$ can be negative, especially when the variable is not very important. In that latter case, the variable is not important enough to make a lot of swings, while correlations with other variables and local over-fitting induce a lot of reverse-swings (i.e adding the variable transforms a winning coalition into a loosing coalition).

3.2 A graphical model: LUCAS

To illustrate our method, we use a dataset generated by the Causal Bayesian network LUCAS[3], used for modeling the occurence of a Lung Cancer, based on a network of 11 binary variables. The variables "Smoking, Coughing, Allergy, Genetics, Fatigue" are Markov Blanket, the 6 other variables are not directly related to the target. We want to explain an observation with a well-defined ground truth. We know from the probability tables that if Smoking, Genetic, Coughing are True, the probability of having Cancer is very high. So, these three variables should have a high Swing-SV. To better analyze the behavior of the Swing-SV values of the new game, they are calculated for different values of π.

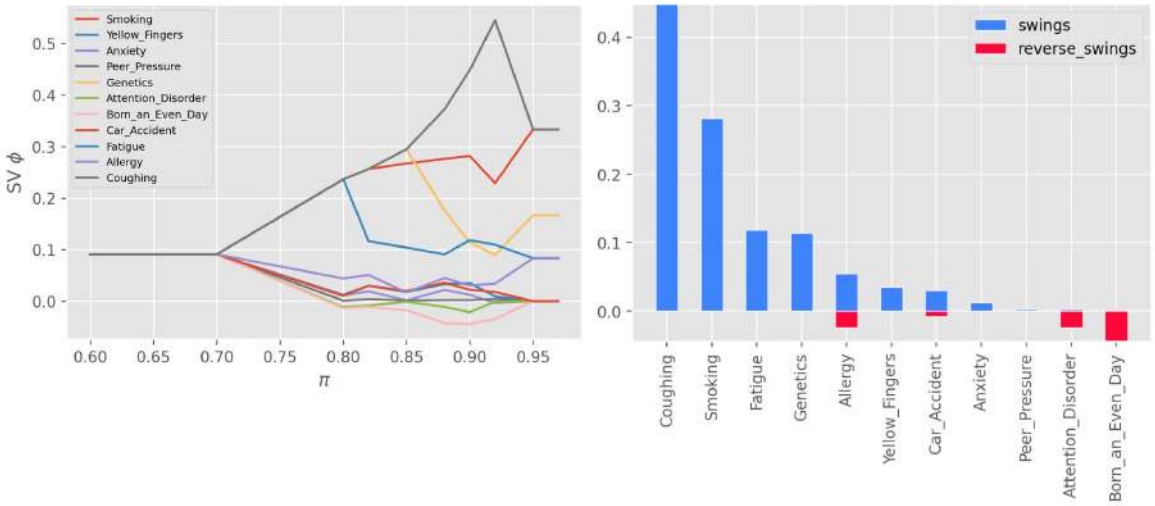


Figure 1: Left: Swing-SV given π. Right: Additive decomposition of the Swing-SV ($\pi = 0.9$).

We observe in the left of figure 1 that all the features have the same Swing-SV for low values of π (below 0.7): all the features have the same rate of swings when the condition is to give the same decision at low level π. For higher probability π, the three expected variables (Smoking, Coughing, Genetic) stand out. The variables Fatigue, Allergy seems important, but the remaining variables have almost zero contributions. In addition, we have also an additive explanation based on the Swing-SV, in order to know if its value comes essentially from the swings or the reverse-swings: we argue that we need to avoid means, as it blurs the interpretation. In the right of figure 1, we remark that important variables do not make any reverse-swings, while irrelevant variables do. Even more, reverse-swings dominate for noisy variables.

[3]http://www.causality.inf.ethz.ch/data/LUCAS.html

4 Conclusion

The Shapley Values for explainable AI are a very useful and insightful methodology for evaluating the importance of variables. While the "standard" Shapley Values can be criticized, we think that the introduction of more adapted game can give a better assessment of the impact of a variable on a decision. In particular, the same decision probability, that evaluates the stability of the decision with respect to fixed subgroups of variables, offers a promising direction for feature attribution and variable selection at a local scale, and possibly at a global scale.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*.

[2] S. Ma and R. Tourani, "Predictive and causal implications of using shapley value for model interpretation," in *Proceedings of the 2020 KDD Workshop on Causal Discovery* (T. D. Le, L. Liu, K. Zhang, E. Kiciman, P. Cui, and A. Hyvärinen, eds.), vol. 127 of *Proceedings of Machine Learning Research*, pp. 23–38, PMLR, 2020.

[3] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable ai: A causal problem," in *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916, PMLR, 2020.

[4] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[5] M. Zaeri-Amirani, F. Afghah, and S. Mousavi, "A feature selection method based on shapley value to false alarm reduction in icus a genetic-algorithm approach," vol. 2018, pp. 319–323, 07 2018.

[6] S. B. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game theory," *Neural Comput.*, vol. 19, no. 7, pp. 1939–1961, 2007.

[7] X. Sun, Y. Liu, J. Li, J. Zhu, X. Liu, and H. Chen, "Using cooperative game theory to optimize the feature selection problem," *Neurocomputing*, vol. 97, pp. 86–93, 2012.

[8] S. Chen, A. Choi, and A. Darwiche, "The same-decision probability: A new tool for decision making," 2012.

Fairness seen as Global Sensitivity Analysis

Clément Bénesse [1], Fabrice Gamboa [2] & Jean-Michel Loubes [2]

[1] *Université Paul Sabatier, Toulouse, clement.benesse@math.univ-toulouse.fr*
[2] *Université Paul Sabatier, Toulouse.*

Résumé. En Fairness, il est important de garantir qu'un prédicteur n'est pas influé par une variable sensible (comme par exemple le genre ou l'origine ethnique). Parallèlement, l'Analyse de Sensibilité fournit un certain nombre d'outils pour quantifier l'influence d'un feature sur la sortie d'un algorithme. Nous réunissons ces deux domaines dans un cadre probabiliste qui permet une définition plus robuste et une compréhension plus fine de la Fairness, avec notamment des applications à l'intersectionalité et aux modèles causaux. Pour cela, nous aurons également besoin d'extensions des indices de Sobol' pour lesquelles nous fournissons un Théorème central limite.

Mots-clés. Fairness, Analyse de Sensibilité, Intersectionalité, Modèles causaux, Indices de Sobol' ...

Abstract. Ensuring that a predictor is not biased against a sensible feature is the key of Fairness learning. Conversely, Global Sensitivity Analysis is used in numerous contexts to monitor the influence of any feature on an output variable. We reconcile these two domains by showing how Fairness can be seen as a special framework of Global Sensitivity Analysis and how various usual indicators are common between these two fields. We also present new Global Sensitivity Analysis indices, as well as rates of convergence, that are useful as fairness proxies.

Keywords. Fairness, Global Sensitivity Analysis, Intersectionality, Causal Models, Sobol' indices ...

1 Introduction

Quantifying the influence of a variable on the outcome of an algorithm is an issue of high importance in order to explain and understand decisions taken by machine learning models. In particular, it enables to detect unwanted biases in the decisions that lead to unfair predictions. This problem has received a growing attention over the last few years in the literature on fair learning for Artificial Intelligence. One of the main difficulty lies in the definition of what is (un)fair and the choices to quantify it. A large number of measures have been designed to assess algorithmic fairness, detecting whether a model depends on variables, called sensitive variables, that convey an information that is irrelevant for the model, from a legal or a moral point of view. We refer for instance to [Dwo+12; Cho17;

OC20] and [BGL20] and references therein for a presentation of different fairness criteria. Most of these definitions stem back to ensuring the independence between a function of an algorithm output and some sensitive feature that may lead to biased treatment. Hence, understanding and measuring the relationships between a sensible feature S, which is typically included in $\mathbf{X}$ or highly correlated to it, and the output of the algorithm $f(\mathbf{X})$ that predicts a target Y, enables to detect unfair algorithmic treatments. Then, ensuring that predictors are fair is achieved by controlling previous measures, as done in [MCE19; WM19; Gra+19; Gor+19; BGL20; Chi+20]. If this notion has been extensively studied for classification, recent work tackle the regression case as in [Gra+19; Jer19; Chz+20] or [LLR20].

Global Sensitivity Analysis (GSA) is used in numerous contexts for quantifying the influence of a set of features on the outcome of a black-box algorithm. Various indicators, usually taking the form of indices between 0 and 1, allow the understanding of how much a feature is important. Multiple set of indices have been proposed over the years such as Sobol' indices, Cramér-von-Mises indices, HSIC – see [JLD06; Da 15; IL15; Gra15; Gam+20] and references therein. The flexibility in the choice allows for deep understanding in the relationship between a feature and the outcome of an algorithm. While a usual assumption in this field is to suppose the inputs to be independent, some works [JLD06; MT12; Gra15] remove this assumption to go further in the understanding of the possible ways for a feature to be influential.
Hence GSA appears to provide a natural framework to understand the impact of sensitive features. This point of view has been considered when using Shapley values in the context of fairness [HSV20] and thus provide local fairness by explainability. Hereafter we provide a full probabilistic framework to use GSA for fairness quantification in machine learning.

Our contribution is two-fold. First, while GSA is usually concerned with independent inputs, we recall extensions of Sobol' indices to non-independent inputs introduced in [MT12] that offer ways to account for joint contribution and correlations between variables while quantifying the influence of a feature. We propose an extension of Cramér-von-Mises indices based on similar ideas. We also prove the asymptotic normality for these extended Sobol' indices to estimate them with a confidence interval. Then, we propose a consistent probabilistic framework to apply GSA's indices to quantify fairness. We illustrate the strength of this approach by showing that it can model classical fairness criteria, causal-based fairness and new notions such as intersectionality. This provides new conceptual and practical perspectives to fairness in Machine Learning.

2 Global Sensitivity Analysis

The main objective of GSA is to monitor the influence of variables $X_1, \cdots, X_p$ on an output variable, or variable of interest, Y. For this, we compare, for a feature X_i and the output Y, the probability distribution $\mathbb{P}_{X_i,Y}$ and the product probability distribution

$\mathbb{P}_{X_i}\mathbb{P}_Y$ by using a measure of dissimilarity. If these two probabilities are equal, the feature X_i has no influence on the output of the algorithm. Otherwise, the influence should be quantifiable. For this, we have access to a wide range of indexes, generally tailored to be valued in $[0, 1]$ and sharing a similar property: the greater the index, the greater the influence of the feature over the outcome. Historically, a variance-decomposition – or Hoeffding decomposition – is used of the output of the black-box algorithm to have access to a second-order moment metric in the so-called Sobol' method. However, these methods were originally developed for independent features. For obvious reasons, this framework is not adapted and has limitations in real-life cases. Additionally, Sobol' methods are intrinsically restrained by the variance-decomposition and others methods have been proposed. We will present two alternatives for Sobol' indices. The first one solves the issue of non-independent features and we provide a Central Limit Theorem for these indices. The second one circumvents the limitations of working with variance-decomposition. We finish by merging these two alternatives, inspired by the works of [AC19; Gam+20] and [Cha20].

3 Fairness

We provide a probabilistic framework to unify all the various Fairness definitions as Global Sensitivity Indices. Several measures of fairness corresponding to different definitions of fairness have been proposed in the machine learning literature. The *Statistical Parity* see for instance in [Dwo+12], requires that the algorithm f, predicting a target Y, has similar outputs for all the values of S in the sense that $\mathbb{P}(f(\mathbf{X}) = 1|S) = \mathbb{P}(f(\mathbf{X}) = 1)$ for general S, continuous or discrete. *Equality of odds* looks for the independence between the error of the algorithm and the protected variable, i.e fairness here implies that $f(\mathbf{X}) \perp\!\!\!\perp S|Y$. This condition is equivalent in the binary case to $\mathbb{P}(f(\mathbf{X}) = 1|Y = i, S) = \mathbb{P}(f(\mathbf{X}) = 1|Y = i)$, for $i = 0, 1$.

Previous notions of fairness are quantified using a *Fairness measure* Λ and a function $\Phi(Y, \mathbf{X})$ such that $\Lambda(\Phi(Y, \mathbf{X}), S) = 0$ in the case of perfect fairness while the constraint is relaxed into $\Lambda(\Phi(Y, \mathbf{X}), S) \leq \varepsilon$, for a small ε, leading to the notion of approximate fairness. The following theorem proves that GSA measures are suitable indicators to quantify fairness as follows and that these definitions can be extended to continuous predictors and continuous Y.

Definition 3.1 *Let Φ be a function of the features $\mathbf{X}$ and of Y. We define a GSA measure for a function Φ and a random variable Z as a $\Gamma(.,.)$ such that $\Gamma(\Phi(Y, \mathbf{X}), Z)$ is equal to 0 if $\Phi(Y, \mathbf{X})$ is independent of Z and is equal to 1 if $\Phi(Y, \mathbf{X})$ is a function of Z.*

Theorem 3.1 *Let Φ be a function of the features and Γ be a GSA measure for Φ and S. Then, Γ induces a Fairness measure defined as $\Lambda(\Phi(Y, \mathbf{X}), S) = \Gamma(\Phi(Y, \mathbf{X}), S)$.*

Table 1: Common fairness definitions and associated GSA measures

Fairness definition	GSA measure associated
Statistical Parity	$\text{Var}(\mathbb{E}[f(\mathbf{X})\|S]) \to Sob_S(f(\mathbf{X}))$
Avoiding Disparate Treatment	$\mathbb{E}[\text{Var}(f(\mathbf{X})\|X)] \to SobT_S(f(\mathbf{X}))$
Equality of odds	$\mathbb{E}[\text{Var}(\mathbb{E}[f(\mathbf{X})\|S,Y]\|Y)] \to CVM^{ind}(f(\mathbf{X}),S\|Y)$
Avoiding Disparate Mistreatment	$\text{Var}(\mathbb{E}[\ell(f(\mathbf{X}),Y)\|S]) \to Sob_S(\ell(f(\mathbf{X}),Y))$

We point out in Table 1 how some well known fairness measures can be linked with GSA indices. We also provide several properties that allow for deeper understanding and extensions of these common fairness measures.

References

[AC19] Mona Azadkia and Sourav Chatterjee. "A simple measure of conditional dependence". In: *arXiv preprint arXiv:1910.12327* (2019).

[BGL20] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. "Review of Mathematical frameworks for Fairness in Machine Learning". In: *arXiv preprint arXiv:2005.13755* (2020).

[Cha20] Sourav Chatterjee. "A new coefficient of correlation". In: *Journal of the American Statistical Association* (2020), pp. 1–21.

[Chi+20] Silvia Chiappa et al. "A General Approach to Fairness with Optimal Transport." In: *AAAI*. 2020, pp. 3633–3640.

[Cho17] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2 (2017), pp. 153–163.

[Chz+20] Evgenii Chzhen et al. "Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees". In: *Advances in Neural Information Processing Systems* (Mar. 2020).

[Da 15] Sébastien Da Veiga. "Global sensitivity analysis with dependence measures". In: *Journal of Statistical Computation and Simulation* 85.7 (May 2015), pp. 1283–1305. DOI: 10.1080/00949655.2014.945932. URL: https://hal.archives-ouvertes.fr/hal-01128666.

[Dwo+12] C. Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM. 2012, pp. 214–226.

[Gam+20] Fabrice Gamboa et al. “Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics”. In: *arXiv preprint arXiv:2003.01772* (2020).

[Gor+19] Paula Gordaliza et al. “Obtaining fairness using optimal transport theory”. In: *International Conference on Machine Learning*. 2019, pp. 2357–2365.

[Gra+19] Vincent Grari et al. *Fairness-Aware Neural Réyni Minimization for Continuous Features*. 2019. eprint: `arXiv:1911.04929`.

[Gra15] Mathilde Grandjacques. “Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes: application en énergétique du bâtiment”. PhD thesis. Grenoble Alpes, 2015.

[HSV20] James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou. *Fairness by Explicability and Adversarial SHAP Learning*. 2020. eprint: `arXiv:2003.05330`.

[IL15] Bertrand Iooss and Paul Lemaître. “A review on global sensitivity analysis methods”. In: *Uncertainty management in simulation-optimization of complex systems*. Springer, 2015, pp. 101–122.

[Jer19] Noureddine El Karoui Jeremie Mary Clement Calauzenes. *Fairness-Aware Learning for Continuous Attributes and Treatments*. 2019.

[JLD06] Julien Jacques, Christian Lavergne, and Nicolas Devictor. “Sensitivity analysis in presence of model uncertainty and correlated inputs”. In: *Reliability Engineering & System Safety* 91.10-11 (2006), pp. 1126–1134.

[LLR20] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. “Projection to fairness in statistical learning”. In: *arXiv e-prints* (2020), arXiv–2005.

[MCE19] Jérémie Mary, Clément Calauzènes, and Noureddine El Karoui. “Fairness-aware learning for continuous attributes and treatments”. In: *International Conference on Machine Learning*. 2019, pp. 4382–4391.

[MT12] Thierry A Mara and Stefano Tarantola. “Variance-based sensitivity indices for models with dependent inputs”. In: *Reliability Engineering & System Safety* 107 (2012), pp. 115–121.

[OC20] Luca Oneto and S Chiappa. *Recent Trends in Learning From Data*. Springer, 2020.

[WM19] Robert C Williamson and Aditya Krishna Menon. “Fairness risk measures”. In: *arXiv preprint arXiv:1901.08665* (2019).

Estimation non-paramétrique dans un modèle de mélange à deux classes

Gaëlle Chagny [1] & Antoine Channarond [1] & Van Hà Hoang [2] & Angelina Roche [3]

[1] *LMRS, UMR CNRS 6085, Université de Rouen Normandie, gaelle.chagny@univ-rouen.fr & antoine.channarond@univ-rouen.fr*
[2] *Faculty of Mathematics et Computer Science, Vietnam National University, Ho Chi Minh City, hvha@hcmus.edu.vn*
[3] *CEREMADE, UMR CNRS 7534, Université Paris Dauphine, roche@ceremade.dauphine.fr*

Résumé. Nous considérons un modèle de mélange de deux lois de probabilité, dont l'une est la loi uniforme sur l'intervalle $[0, 1]$, et on s'intéresse à l'estimation non-paramétrique et adaptative de la densité de probabilité de la seconde composante du mélange. Ce problème apparait par exemple dans des questions d'estimation robuste et dans les procédures de contrôle du taux de faux positifs dans un contexte de tests multiples. Nous définissons un estimateur à noyau pondéré, à sélection de fenêtre automatique, selon une méthode inspirée de Goldenshluger et Lepski (2011). Sa construction implique l'introduction de contreparties empiriques, à la fois pour la densité mélange et pour la proportion de chaque classe du mélange : des estimateurs préliminaires pour ces deux quantités sont également proposés. Une inégalité de type oracle est obtenue pour le risque ponctuel, et la vitesse de convergence est calculée lorsque la fonction à estimer est suffisamment régulière. Ces résultats théoriques sont illustrés par des simulations numériques.

Mots-clés. Estimation non-paramétrique adaptative, modèles de mélange, estimateurs à noyaux, sélection de fenêtre.

Abstract. We consider a two-class mixture model, where the density of one of the components is known (equal to the uniform density on the interval $[0; 1]$). This problem appears in some statistical settings, robust estimation and multiple testing among others. We address the issue of the nonparametric adaptive estimation of the unknown probability density of the second component. We propose a randomly weighted kernel estimator with a fully data-driven bandwidth selection method, in the spirit of Goldenshluger and Lepski (2011). Its definition involves empirical counterparts both for the mixture density and the mixing proportion : preliminary estimators for these quantities are also proposed. An oracle-type inequality for the pointwise quadratic risk is derived as well as convergence rates over Hölder smoothness classes. The theoretical results are illustrated by numerical simulations.

Keywords. Adaptive non-parametric estimation, mixture models, kernel estimator, bandwidth selection.

1 Introduction

Nous considérons un modèle de mélange à deux classes, de la forme

$$g(x) = \theta + (1-\theta)f(x), \quad \forall x \in [0,1], \tag{1}$$

et dans lequel à la fois la proportion du mélange $\theta \in (0,1)$ et la densité de probabilité f (par rapport à la mesure de Lebesgue sur $[0;1]$) sont inconnues. À partir d'un échantillon de variables aléatoires $X_1, \ldots, X_n$, $n \in \mathbb{N}\backslash\{0\}$ indépendantes et identiquement distribuées (*i.i.d.* dans la suite) de densité g (inconnue elle aussi), nous souhaitons principalement construire une stratégie d'estimation adaptative optimale pour la densité composante f. L'estimation de la proportion θ, qui apparaît comme une étape intermédiaire dans ce programme, est également traitée.

Le modèle (1) apparaît dans divers contextes statistiques. On peut tout d'abord le voir comme un modèle où la densité cible f est contaminée, en une certaine proportion θ, par la loi uniforme sur $[0,1]$: l'objectif est alors d'estimer f de manière robuste à partir des observations contaminées $X_1, \ldots, X_n$. Cependant, en ce sens, la proportion de contamination θ est généralement supposée connue (alors que la distribution contaminante ne l'est pas toujours), et les résultats théoriques existants proposent des vitesses de convergence minimax comme fonctions de n et de θ (voir Liu et Gao, 2017). Dans le cadre des tests multiples, si l'on suppose que l'on teste un grand nombre n d'hypothèses, de manière indépendante, et simultanément, alors les $p-$valeurs $X_1, \ldots, X_n$ générées par ces tests peuvent être modélisées par (1). En effet, sous l'hypothèse nulle, elles suivent la loi uniforme sur $[0,1]$, et ont une densité inconnue f sous l'hypothèse alternative. Le paramètre θ représente la proportion asymptotique de vraies hypothèses nulles. L'estimation de f peut-être requise par exemple pour évaluer et contrôler différents types d'erreurs de la procédure de tests, comme le "False discrovery rate", voir par exemple Langaas *et al.* (2005), Robin *et al.* (2007), Nguyen et Matias (2014a,b).

L'estimation de f dans le modèle (1) a donc déjà été abordée dans la littérature. Langaas *et al.* (2005) ont par exemple proposé un estimateur basé sur une vraisemblance non-paramétrique, mais les caractéristiques théoriques de la méthode n'ont pas été étudiées. Robin *et al.* (2007) ont construit un estimateur à noyau pondéré, dont les poids sont des estimateurs des probabilités *a posteriori* du mélange (c'est à dire les probabilités pour chaque individu i d'appartenir à la composante non-paramétrique). Un algorithme itératif de type EM est proposé, et les auteurs montrent qu'il converge vers une unique solution. L'étude asymptotique de cet estimateur à noyau est reprise par Nguyen et Matias (2014a), qui calculent une vitesse de convergence pour le risque ponctuel, sur des classes de Hölder. Cependant, la procédure d'estimation ne s'adapte pas à la régularité, supposée dans cet article connue, de la fonction inconnue f.

Nous proposons ici une stratégie d'inférence complète à la fois pour f et θ. Un nouvel estimateur à noyau est construit, et une sélection automatique de la fenêtre est pro-

posée. Nous démontrons des résultats théoriques pour la procédure d'estimation complète (inégalité de type oracle, vitesse de convergence pour le risque quadratique ponctuel) qui prouvent le caractère adaptatif et optimal au sens minimax de notre méthode.

2 Collection d'estimateurs à noyaux pour la densité cible

La principale difficulté pour construire des estimateurs de f dans le modèle (1) provient du fait que nous ne disposons pas d'observations directement tirées selon la densité f, mais d'observations tirées selon g. Un estimateur à noyau classique de la densité ne peut donc pas être proposé. La clé de notre approche est l'égalité suivante, qui relie la densité inconnue f à la loi g des variables observées : pour tous $\theta, x \in [0;1]$,

$$f(x) = w(\theta, g(x))g(x), \text{ avec } w(\theta, g(x)) := \frac{1}{1-\theta}\left(1 - \frac{\theta}{g(x)}\right). \tag{2}$$

Nous utilisons cette relation comme suit. Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau, c'est-à-dire une fonction intégrable telle que $\int_{\mathbb{R}} K(x)dx = 1$ et $\int_{\mathbb{R}} K^2(x)dx < +\infty$. Pour tout $h > 0$, soit $K_h(\cdot) = K(\cdot/h)/h$. Alors, si Y est une variable de densité g, $\mathbb{E}[K_h(x - Y)] = \mathbb{E}[w(\theta, g(X_1))K_h(x - X_1)]$. Ceci conduit à la définition

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w(\tilde{\theta}_n, \hat{g}(X_i))K_h(x - X_i), \text{ avec } w(\tilde{\theta}_n, \hat{g}(X_i)) = \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{\hat{g}(X_i)}\right), \tag{3}$$

où $\tilde{\theta}_n$ et $\hat{g}$ sont des estimateurs préliminaires pour θ et g respectivement, définis à partir d'un échantillon additionnel $(X_i)_{i=n+1,\dots,2n}$ de variables *i.i.d.* selon la densité g, indépendant de $(X_i)_{i=1,\dots,n}$. En comparaison avec l'estimateur à noyau pondéré de Robin *et al.* (2007) et Nguyen et Matias (2014a), le principal avantage de notre méthode est qu'elle aboutit à $\mathbb{E}[\hat{f}_h(x)] = K_h \star f(x)$, où $\star$ désigne le produit de convolution, dans le cas où $\hat{g}$ et $\tilde{\theta}_n$ sont remplacés par leurs équivalents théoriques g et θ dans (3). Une telle égalité est cruciale pour l'étude du biais de l'estimateur, et donc pour mettre en oeuvre une méthode automatique de sélection de fenêtre.

Sous des hypothèses concernant le choix du noyau K et de la fenêtre h, la densité f (qui doit être bornée sur un voisinage $\mathcal{V}_n(x_0)$ du point d'estimation x_0), et les estimateurs préliminaires $\tilde{\theta}_n$ et $\hat{g}$ (qui doivent être suffisamment précis), nous démontrons la borne supérieure suivante: il existe des constantes C_ℓ, $\ell = 1, \dots, 4$, telles que

$$\mathbb{E}\left[\left(\widehat{f}_h(x_0) - f(x_0)\right)^2\right] \leq C_1 \left\{(K_h \star f(x_0) - f(x_0))^2 + \frac{1}{\gamma^2 nh}\right\} \tag{4}$$
$$+ C_2\mathbb{E}\left[\left|\tilde{\theta}_n - \theta\right|^2\right] + C_3\mathbb{E}\left[\|\hat{g} - g\|^2_{\infty,\mathcal{V}_n(x_0)}\right] + \frac{C_4}{n^2}$$

La notation γ désigne l'infimum de g sur le voisinage $\mathcal{V}_n(x_0)$. Il s'agit d'une décomposition biais-variance du risque : le premier terme du membre de droite de (4) est un terme de biais qui décroît quand la fenêtre h tend vers 0, alors que le second, qui est le terme de variance, croit quand $h \to 0$. Les termes de la ligne suivante sont inévitables, et dûs au plug-in des estimateurs préliminaires $\tilde{\theta}_n$ et $\hat{g}$. Nous montrons plus loin qu'ils ne dégradent pas la vitesse de convergence.

3 Sélection automatique de la fenêtre

Considérant une collection finie $\mathcal{H}_n$ (de cardinal borné par n), nous souhaitons ensuite sélectionner le *meilleur* estimateur dans la collection $(\widehat{f}_h)_{h\in\mathcal{H}_n}$: celui qui a la plus petite décomposition biais-variance possible, et donc qui minimise, sur $\mathcal{H}_n$, le membre de droite de (4). En pratique, ces termes impliquent des quantités inconnues, car dépendant de la fonction cible f. Nous proposons donc un critère ponctuel fondé sur les données uniquement, dans l'esprit de la méthode de Goldenshluger et Lepski (2011).

On sélectionne $\widehat{h}(x_0) = \arg\min_{h\in\mathcal{H}_n}\{A(h,x_0) + V(h,x_0)\}$ où V approche le terme de variance du risque et A estime le terme de biais :

$$V(x_0,h) = \frac{\kappa\|K\|_1^2\|K\|_2^2\|g\|_{\infty,V_n(x_0)}\log(n)}{\hat{\gamma}^2 nh}$$

pour une constante $\kappa > 0$ à calibrer en pratique, $\hat{\gamma}$ un estimateur de γ et

$$A(x_0,h) = \max_{h'\in\mathcal{H}_n}\left\{\left(\widehat{f}_{h,h'}(x_0) - \widehat{f}_{h'}(x_0)\right)^2 - V(x_0,h')\right\}_+,$$

avec, pour $h,h' \in \mathcal{H}_n$, $\widehat{f}_{h,h'}(x_0) = (K_{h'}\star\widehat{f}_h)(x_0) = n^{-1}\sum_{i=1}^n w(\tilde{\theta}_n, \hat{g}(X_i))(K_h\star K_{h'})(x_0 - X_i)$ les estimateurs auxiliaires spécifiques de la méthode. Sous des conditions détaillées dans le preprint Chagny *et al.* (2020), nous démontrons l'inégalité de type oracle suivante, qui prouve que l'estimateur sélectionné par la méthode ci-dessus fait aussi bien, à constante multiplicative près, et à termes de reste près, que le meilleur des estimateurs de la collection :

Théorème 1. *L'estimateur sélectionné* $\widehat{f}(x_0) := \widehat{f}_{\widehat{h}(x_0)}(x_0)$ *vérifie, pour* $\delta \in (0,1)$*, et pour des constantes* C_ℓ*,* $\ell = 5, \ldots, 8$*.*

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{f}(x_0) - f(x_0)\right)^2\right] \;\leq\; & C_5 \min_{h\in\mathcal{H}_n}\left\{\|K_h \star f - f\|^2_{\infty,V_n(x_0)} + \frac{\log(n)}{\gamma^2 nh}\right\} \\ & + C_6 \sup_{\theta\in[\delta,1-\delta]} \mathbb{E}\left[\left|\tilde{\theta}_n - \theta\right|^2\right] + C_7\mathbb{E}\left[\|\hat{g} - g\|^2_{\infty,V_n(x_0)}\right] + \frac{C_8}{n^2},\end{aligned}$$

4 Estimation préliminaire pour la densité mélange et la proportion de chaque classe du mélange

Cette section est consacrée à de brèves explications concernant la construction choisie pour les estimateurs préliminaires $\hat{g}$ et $\tilde{\theta}_n$ à partir de l'échantillon supplémentaire $(X_i)_{i=n+1,\ldots,2n}$ tiré selon g, de telle sorte que ceux-ci satisfassent les hypothèses dont nous avons besoin pour démontrer les résultats ci-dessus, et atteignent une vitesse de convergence permettant d'obtenir l'optimalité de notre procédure d'estimation de f.

Pour estimer g, nous nous inspirons du travail de Bertin *et al.* (2016), dont l'estimateur de la densité conditionnelle requiert également un estimateur de densité préliminaire. Nous utilisons un estimateur à noyau classique, de la forme $\hat{g}_{\hat{b}}(x_0) = n^{-1}\sum_{i=n+1}^{2n} L_{\hat{b}}(x_0 - X_i)$, pour $L_b(x) = L(x/b)/b$, $b > 0$, et L un noyau, et où $\hat{b}$ est sélectionnée avec une méthode de Lepski.

Pour estimer θ, de nombreuses méthodes ont été proposées dans la littérature. Nous nous restreignons à un cas particulier du modèle (1), qui assure l'identifiabilité du couple (θ, f). Nous supposons que la densité f appartient à l'ensemble

$$\mathcal{F}_\delta = \left\{f : [0,1] \to \mathbb{R}_+, f \text{ est une densité continue telle que } f_{|[1-\delta,1]} = 0\right\}$$

pour un $\delta > 0$. Ainsi, $\theta = \inf_{x\in[0,1]} g(x) = g(1)$, et une idée naturelle consiste à insérer un estimateur de g pour retrouver θ. Pour être cohérent avec le restant de la procédure, nous utilisons un estimateur à noyau pour g sur $[0,1]$. Mais, pour éviter les effets de bord inhérents à ce type de méthode (qui conduiraient à un mauvais estimateur de θ), nous utilisons un principe de réflexion, qui conduit à un estimateur à noyau $\hat{g}^{sym}_{\hat{b}}$ sur l'intervalle $[0,2]$, où la fenêtre $\hat{b}$ est sélectionnée automatiquement par méthode de Goldenshluger-Lepski. Enfin, nous posons

$$\hat{\theta}_{n,\hat{b}} = \frac{1}{\delta}\int_{1-\delta}^{1+\delta} \hat{g}^{sym}_{\hat{b}}(x)dx,$$

et $\tilde{\theta}_{n,\hat{b}} := \max(\min(\hat{\theta}_{n,\hat{b}}, 1-\delta/2), \delta/2)$, pour assurer l'une de nos hypothèses, $\tilde{\theta}_{n,\hat{b}} \in [\delta/2, 1-\delta/2]$.

Ainsi, nous démontrons qu'avec ces estimateurs préliminaires, l'estimateur de f en x_0 défini en (3) avec fenêtre sélectionnée par méthode de Goldenshluger-Lepski converge à la vitesse minimax attendue, tout en s'adaptant automatiquement à la régularité inconnue de f, à un terme en log prêt (inévitable en estimation ponctuelle) :

Corollaire 2. *En utilisant les estimateurs préliminaires définis ci-dessus, si f appartient à une boule d'un espace de Hölder de régularité $\beta > 0$, et si la collection de fenêtres et le noyau sont bien choisis, alors*

$$\mathbb{E}\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] \leq C_9 \left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}},$$

pour une constante $C_9 > 0$.

Des détails et commentaires concernant la procédure, ainsi qu'une application sur des données simulées, peuvent être trouvés dans le preprint de Chagny *et al.* (2020).

Bibliographie

Bertin, K., Lacour C., et Rivoirard, V. (2016) Adaptive pointwise estimation of conditional density function. *Annales de l'Institut Henti Poincaré, Probabilités et Statistiques*, 52(2), pp. 939-980.

Chagny G., Channarond. A., Hoang V-H., et Roche A. (2020) Adaptive nonparametric estimation of a component density in a two-class mixture model, *Preprint, hal-02909601.*

Goldenshluger A. et Lepski, O. (2011) Betwidth selection in kernel density estimation: oracle inequalities et adaptive minimax optimality. *The Annals of Statistics*, 39(3), pp. 1608-1632.

Langaas, M. Lindqvist, B-H, et Ferkingstad E. (2005) Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4), pp.555-572.

Liu H. et Gao C. (2017) Density estimation with contaminated data: Minimax rates et theory of adaptation. *Preprint, arXiv:1712.07801.*

Nguyen, V-H. et Matias, C.(2014a) Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. *ESAIM: Probability et Statistics*, 18, pp. 584-612.

Nguyen, V-H. et Matias, C. (2014b). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scetinavian Journal of Statistics*, 41(4), pp. 1167-1194.

Robin, S., Bar-Hen, A. Daudin, J.J. et Pierre, L. (2007) A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12), pp. 5483-5493..

Nonstationary Nearest Neighbor Gaussian Process : hierarchical model architecture and MCMC sampling

Sébastien Coube-Sisqueille [1] & Sudipto Banerjee [2] & Benoît Liquet [3]

[1] *sebastien.coube@univ-pau.fr*
[2] *sudipto@ucla.edu*
[3] *benoit.liquet@univ-pau.fr*

Résumé. La modélisation spatiale non stationnaire est une approche intéressante et prometteuse, mais elle souffre de plusieurs problèmes : son coût computationel, la complexité et le manque de lisibilité de modèles hiérarchiques à plusieurs étages, et la difficulté de sélectionner un modèle. Nous répondons à ces trois problèmes en introduisant un modèle non stationnaire utilisant les processus gaussiens des plus proches voisins (NNGP, pour *Nearest Neighbor Gaussian Process*).
Les NNGP, précis et économiques, sont un bon départ pour répondre au problème du temps de calcul. Nous étudions le comportement des NNGP utilisant une fonction de covariance non stationnaire, d'une part en déduisant des propriétés analytiques et d'autre part en testant empiriquement l'impact de l'ordre des observations sur la covariance qui est induite par un NNGP.
Nous introduisons une architecture de modèle lisible afin de faciliter la compréhension des résultats et la sélection de modèles. En particulier, nous créons une famille de modèles cohérente qui rassemble les processus spatiaux avec portée stationnaire, les processus non stationnaires avec des paramètres de portée circulaires, et ceux avec des paramètres de portée elliptiques.
Nous tirons parti de notre architecture hiérarchique et de l'utilisation des NNGP en proposant deux algorithmes MCMC *ad hoc*, basés respectivement sur un algorithme de Langevin ajusté par un pas de Metropolis et sur un échantillonneur chromatique. Nous améliorons ces deux algorithmes en utilisant la méthode de l'entremêlement de paramétrisations.
Nous testons nos méthodes avec des jeux de données synthétiques pour trouver des règles empiriques concernant le choix de l'algorithme MCMC, des hyperparamètres, ainsi que la sélection de modèle. Nous les utilisons pour analyser un jeu de donnés de pollution au plomb aux États-Unis d'Amérique.

Mots-clés. Processus gaussien des Plus Proches Voisins, Modèle spatial non stationnaire, MCMC

Abstract. Nonstationary spatial modelling is exciting and potentially rewarding, but suffers from several problems : its computational cost, the complexity and lack of interpretability of multi-layered hierarchical models, and the difficulty of model selection. We

tackle those problems by introducing a nonstationary Nearest Neighbor Gaussian Process (NNGP) model.
NNGPs are a good starting point to address the problem of the computational cost because of their accuracy and affordability. We study the behavior of NNGPs that use a nonstationary covariance function, deriving some algebraic properties and exploring the impact of ordering on the effective covariance induced by NNGPs.
To ease results analysis and model selection, we introduce a readable hierarchical model architecture. In particular, we make parameters interpretation and model selection easier by integrating stationary range, nonstationary range with circular parameters, and nonstationary range with elliptic parameters in a consistent framework.
Given the NNGP approximation and the model architecture, we propose two *ad hoc* MCMC algorithms based on Metropolis Adjusted Langevin Algorithm and Chromatic Sampling, both being improved using interweaving of parametrizations.
We carry out experiments on synthetic data sets to find empirical practical rules concerning on MCMC algorithm choice, hyperparameter tuning, and model selection. Finally, we use those guidelines to analyze a data set of lead contamination in the United States of America.
Keywords. Nearest Neighbor Gaussian Process, Nonstationary Spatial Modelling, MCMC

1 Modèle spatial non stationnaire

Nous observons une variable d'intérêt sur une collection de sites spatiaux $\mathcal{S}$. Nous partons d'une modélisation spatiale stationnaire et considérons trois extensions où différents paramètres peuvent varier dans l'espace : la variance marginale du processus latent $\sigma^2(\mathcal{S})$, les paramètres de portée (potentiellement elliptiques) du processus latent, et, quand les observations sont gaussiennes, la variance du bruit $\tau^2(\mathcal{S})$. Ces trois augmentations sont mutuellement compatibles et sont résumées dans les équations suivantes.
Les observations gaussiennes sont analysées comme :

$$z(s) = X(s)\beta^T + w(s) + \epsilon(s).$$

Le bruit gaussien est paramétrisé :

$$\epsilon(\mathcal{S}) \sim \mathcal{N}\left(0,\ diag(\tau^2(s_1) \dots \tau^2(s_n))\right) \tag{1}$$

où $\tau(s_1 \dots s_n)$ est une collection d'écarts types qui varient dans l'espace. Le champ latent est paramétrisé comme une loi normale multivariée

$$w(\mathcal{S}) \sim \mathcal{N}(0, \Sigma).$$

La fonction de covariance entre deux sites spatiaux est donnée par

$$\Sigma_{i,j} = K(s_i, s_j) = \sigma(s_i)\sigma(s_j)K_0(s_i, s_j, \alpha(s_i), \alpha(s_j)) \tag{2}$$

où $\sigma(s_1 \dots s_n)$ est une collection d'écarts types qui varient dans l'espace, K_0 st une fonction de corrélation, et $\alpha(s_1, \dots, s_n)$ est une collection de paramètres de portée. Ces paramètres peuvent être des matrices définies positives donnant une covariance localement anisotropique, ou des nombres positifs donnant une covariance localement isotropique. Le premier cas est donné par Paciorek (2003).

$$K_0(s, t, A(s), A(t)) = \frac{2^{d/2}|A(s)|^{1/4}|A(t)|^{1/4}}{|A(s) + A(t)|^{1/2}} K_i\left(d_M\left(s, t, (A(s) + A(t))/2\right)\right) \tag{3}$$

$A(s)$ et $A(t)$ étant des matrices de portée, d étant la dimension du domaine spatial ou spatio-temporel, $d_M(\cdot, \cdot, \cdot)$ étant la distance de Mahalanobis, et K_i étant une fonction de corrélation stationnaire. Si toutes les matrices sont égales, alors la corrélation est stationnaire. On peut construire une fonction localement isotropique en contraignant $A = \alpha I_d$ et en identifiant la distance de Mahalanobis utilisant une matrice diagonale avec la distance Euclidienne $d_E(\cdot, \cdot)$:

$$K_0(s, t, \alpha(s), \alpha(t)) = \left(\frac{\sqrt{2}\alpha(s)^{1/4}\alpha(t)^{1/4}}{(\alpha(s) + \alpha(t))^{1/2}}\right)^d K_i\left(d_E(s, t)/\left((\alpha(s) + \alpha(t))/2\right)\right). \tag{4}$$

2 Modélisation non stationnaire utilisant un processus gaussiens des plus proches voisins

2.1 Présentation des Processus gaussiens des plus proches voisins

Le calcul de la densité normale multivariée du champ latent $w(\mathcal{S})$ implique l'inversion et le calcul du déterminant de la matrice de covariance. Cette méthode devient inabordable quand il y a plus de quelques milliers d'observations. Les Processus gaussiens des Plus Proches Voisins (NNGP) et plus généralement la famille des Approximations de Vecchia ont prouvé au cours des dernières années qu'elles permettent d'approximer à faible coût les densités gaussiennes (Vecchia, 1988; Stein, Chi & Welty, 2004; Datta, Banerjee, Finley & Gelfand, 2016; Finley, Datta, Cook, Morton, Andersen & Banerjee, 2019; Katzfuss & Guinness, 2017; Guinness, 2018). La densité NNGP (stationnaire) est définie en utilisant une densité conditionnelle récurrente "élaguée" :

$$\tilde{f}(w(s_i)|w(s_1, \dots, s_{i-1}), \theta) = f(w(s_i)|w(pa(s_i)), \theta), \tag{5}$$

$pa(s_i)$ étant les parents du site s_i dans un Graphe Dirigé Acyclique (DAG) dont les sommets sont identifiés avec les observations, $\tilde{f}(\cdot)$ la densité NNGP, et $f(\cdot)$ étant la

densité gaussienne non approximée avec des paramètres de covariance θ. L'ordre des sites spatiaux utilisé pour définir le DAG est déterminant pour la qualité de l'approximation (Guinness, 2018). L'heuristique de choix des parents la plus populaire est de prendre les plus proches voisins spatiaux parmi les points qui précèdent dans le DAG, donnant le nom de Processus gaussien des Plus Proches voisins.
Les NNGP ont la propriété très désirable de permettre d'obtenir facilement un facteur de Cholesky de la matrice de précision avec très peu de coefficients non-nuls. Nous notons ce facteur $\tilde{R}$ (la matrice de covariance induite est alors $\left(\tilde{R}^T\tilde{R}\right)^{-1}$).

2.2 Propriétés des Processus gaussiens des plus proches voisins avec une fonction de covariance non stationnaire

Un package R (Guinness, 2018) comprend des fonctions de covariance non-stationnaires, mais à notre connaissance il n'y a pas eu à ce jour d'étude théorique ou empirique des propriétés des NNGP quand on utilise une fonction de covariance non stationnaire.
Un premier aspect de notre travail a été de préciser des propriétés algébriques des NNGP utilisant une fonction de covariance donnée par (2). La densité n'est plus paramétrisée par un unique jeu de paramètres de covariance θ mais par des paramètres qui varient dans l'espace $\theta(\mathcal{S})$. Nous avons réussi à réécrire (5) sous la forme

$$\tilde{f}(w(s_i)|w(s_1,\ldots,s_{i-1}),\theta(\mathcal{S})) = f(w(s_i)|w(pa(s_i)),\theta(s_i \cup pa(s_i))). \tag{6}$$

Cela veut dire qu'au lieu de conditionner par tous les paramètres de covariance il est possible de conditionner seulement par les paramètres des parents dans le DAG. Ce résultat permet d'utiliser les puissantes propriétés de Markov qui découlent des factorisations sur un graphe (Lauritzen, 1996).

L'autre propriété concerne la variance marginale. Notons $\tilde{R}_0$ le facteur de Cholesky NNGP obtenu en utilisant la corrélation $K_0(\cdot)$ des équations (3) ou (4). On peut alors réécrire le facteur de Cholesky NNGP sous la forme

$$\tilde{R} = \tilde{R}_0 \; diag(\sigma(\mathcal{S}))^{-1}. \tag{7}$$

Comme pour un processus gaussien non-approximé, il est possible de factoriser la matrice de covariance d'un NNGP. En pratique, cela permet de calculer la densité marginale de $\sigma(\mathcal{S})$ à faible coût.

En pratique, toutes les approximations NNGP ne se valent pas. Guinness (2018) montre, pour des processus stationnaires, que certains agencements des points lors de la construction du DAG donnent de bien meilleurs résultats. Nous confirmons ces observations dans le cas de processus non stationnaires. Les méthodes repérées par Guinness donnent des résultats satisfaisants, alors que les autres méthodes provoquent l'apparition d'artefacts.
Nous constatons que la qualité des NNGP se dégrade dans le cas de covariances extrêmement

anisotropiques, et ce même dans le cas stationnaire. L'approximation NNGP est donc réservée à des données exhibant une anisotropie modérée, ce qui laisse cependant un large spectre d'applications.

Il n'existe pas d'équivalent immédiat de (3) et (4) sur la sphère. Nous définissons un NNGP sur la sphère en calculant chaque composante de (5) sur le plan tangent en s_i. Les sites parents sont envoyés dans ce plan par projection orthogonale. Cela nous éloigne d'une vision des NNGP comme approximation à un processus gaussien puisque nous définissons un NNGP sur la sphère sans connaître de fonction de covariance associée.

3 Architecture et paramétrisation du modèle non stationnaire

Dans le but d'imposer une cohérence spatiale ou spatio-temporelle à un champ de paramètres, nous utilisons un processus log-gaussien (log-GP) comme Heinonen, Mannerström, Rousu, Kaski & Lähdesmäki (2016). Le champ latent $\theta(\mathcal{S})$ est analysé comme :

$$log(\theta(s)) = w_\theta(s) + X_\theta(s)\beta_\theta^T \quad \forall\ s \in \mathcal{S} \quad \text{and} \quad w_\theta(\mathcal{S}) \sim \mathcal{N}(0, \theta_\theta). \tag{8}$$

Le processus gaussien $w_\theta(\cdot)$ permet de modéliser des variations ayant une cohérence spatiale. Les coefficients de régression linéaires β_θ paramétrisent des effets fixes (entre autres un intercept), et θ_θ est un jeu de paramètres de covariance qui paramétrisent le prior log-GP.
Les effets fixes permettent de lier le champ de paramètres à des variables environnementales d'intérêt (il est connu, par exemple, que l'altitude rend les précipitations imprévisibles) ou à des variables dépendant des coordonnées spatiales qui vont compléter un processus gaussien pour capturer des variations spatiales. La parmétrisation logarithmique est facile à interpréter, ce qui rend l'utilisation du modèle plus intuitive. Premièrement, les paramètres de covariance telles qu'une portée ou une variance sont des nombres positifs alors qu'un processus gaussien peut prendre toutes les valeurs. Prendre les logarithmes de ces paramètres garantit la validité du prior. Dans le prolongement de cet argument, les paramètres de covariance sont des tailles : une variance est la taille d'une distribution, une portée est la largeur d'un kernel. Il est plus naturel d'utiliser le logarithme pour les comparer car celui-ci plonge l'échelle des ratios dans l'échelle des intervalles (Stevens *et al.*, 1946). Enfin, certains problèmes de paramétrisation disparaissent d'eux-mêmes dans le cadre de ce modèle. Par exemple, on peut se demander s'il vaut mieux utiliser une variance, une précision, ou un écart type pour paramétriser la variance du bruit ou la variance marginale du champ latent. Une fois passées au logarithme, ces paramétrisations diffèrent uniquement par une constante multiplicative. Le problème de paramétrisation devient un problème de paramètres, qui peut être résolu par estimation ou réglage.

Cela ne règle pas le problème des covariances définies par (3), dont les paramètres de portée sont des matrices définies positives. Nous introduisons un prior original utilisant

le logarithme matriciel. Rappelons que le logarithme d'une matrice définie positive est obtenu en passant les valeurs propres au logarithme, et qu'il est bijectif entre les matrices définies positives et les matrices symétriques. De façon similaire à (8), nous analysons le logarithme des matrices de portée de la façon suivante :

$$log(A(s)) = W(s) + \Sigma_{i=i}^{n_{XA}} X_i(s) \times B_i, \tag{9}$$

n_{XA} étant le nombre de variables, $X_i(\cdot)\ 1 \leq i \leq n_{XA}$, la $i^{ème}$ variable, et $B_i, 1 \leq i \leq n_{XA}$ étant une matrice symétrique de taille $d \times d$. Vu que B_i ne dépende pas de s, $\Sigma_{i=i}^{k} X_i(s) \times B_i$ s'interprète comme un effet fixe.
L'effet aléatoire de cette formule est $W(s)$, une matrice symétrique qui dépend s. De façon analogue à (8), nous définissons une distribution dans l'espace des matrices $\Psi(\cdot|\theta_A)$

$$\begin{aligned} &\Psi(W(\mathcal{S})|\theta_A) = \left((2\pi)^{-n/2}|\Sigma(\mathcal{S},\theta_A)|^{-1/2}\right)^{d(d+1)/2} \times \\ &\exp\left(-1/2\Sigma_{i\in1,\ldots,n}\Sigma_{j\in1,\ldots,n}(\Sigma(\mathcal{S},\theta_A)^{-1})_{i,j}\left\langle W(s_i), W(s_j)\right\rangle_F\right) \end{aligned} \tag{10}$$

avec $\langle\cdot,\cdot\rangle_F$ qui est le produit de Frobenius, et $\Sigma(\mathcal{S},\theta_A)$ une matrice de covariance de taille $n \times n$ qui dépende de paramètres de covariance θ_A. Notons que Ψ ressemble à une distribution normale où la multiplication scalaire serait remplacée par le produit de Frobenius. Nous montrons que cette distribution est non seulement valide mais également facile à manipuler car les projections de $W(\cdot)$ sur une base orthonormale quelconque des matrices symétriques suivent des distributions gaussiennes indépendantes.

Afin de pouvoir passer à l'échelle, nous approximons le processus log-gaussien et son extension matricielle en utilisant un NNGP au lieu d'un processus gaussien complet. Parmi les paramètres de covariance du processus log-gaussien, nous estimons la variance marginale mais pas la portée, que nous traitons comme un hyperparamètre.

4 Stratégies MCMC

Une grande partie du travail de cet article a été de trouver des stratégies MCMC adaptées à de grands jeux de données spatiales. En effet, les modèles non stationnaires que nous avons étudiés sont appliqués à des données comptant au plus quelques centaines (Fuglstad, Lindgren, Simpson & Rue, 2015*a*; Heinonen *et al.*, 2016) ou milliers d'observations (Fuglstad, Simpson, Lindgren & Rue, 2015*b*). La littérature présente plusieurs stratégies MCMC destinés à des modèles stationnaires telles que Datta *et al.* (2016); Finley *et al.* (2019); Coube & Liquet (2020). Cependant, la partie non-stationnaire de notre modèle étant nouvelle, nous avons du développer des algorithmes *ad hoc*. Nous avons trouvé deux approches.

La première approche est un échantillonage chromatique (Gonzalez, Low, Gretton & Guestrin, 2011). Utilisant d'une part la propriété de factorisation des NNGP non stationnaires (6), et d'autre part l'indépendance conditionnelle induite par l'utilisation

d'une approximation NNGP au prior log-gaussien, nous avons prouvé qu'un échantilloneur chromatique peut être utilisé pour les paramètres de portée et de variance marginale du processus latent. Un résultat analogue est démontré pour la variance du bruit gaussien hétéroscédastique (1). En pratique, cette méthode est applicable pour la variance du bruit gaussien, pour la variance du processus latent grâce à la propriété de factorisation des NNGP (7), mais pas aux paramètres de portée en raison du grand nombre de calculs de l'approximation NNGP qui seraient nécessaires.

La seconde approche est un algorithme de Langevin ajusté par un pas de Metropolis inspiré du Monte-Carlo hybride de Heinonen *et al.* (2016). Dans le cas des variances du bruit et du processus latent, ce pas est simple à implémenter, encore une fois grâce à (7) dans le second cas. Pour les paramètres de portée, la méthode a besoin du gradient du facteur de Cholesky NNGP $\tilde{R}$ par rapport aux paramètres de portée. Ce facteur de Cholesky est défini ligne par ligne avec (6). Bien que la formule du gradient soit fastidieuse, elle peut être implémentée relativement efficacement et permet donc d'utiliser l'algorithme de Langevin pour échantillonner les paramètres de portée. Nous remarquons aussi que le gradient que nous avons obtenu pourrait servir à d'autres méthodes, telles que l'approche de maximum de vraisemblance développée par Guinness (2018).

Dans les deux cas, nous utilisons les méthodes d'entremêlement de paramétrisation développées par Yu & Meng (2011) et Filippone, Zhong & Girolami (2013) afin d'améliorer le comportement des chaînes MCMC.

5 Références bibliographiques

Notre travail s'inscrit dans les travaux sur les NNGP (Datta *et al.*, 2016; Finley *et al.*, 2019; Coube and Liquet, 2020) et plus largement les approximations de Vecchia (Vecchia, 1988; Stein *et al.*, 2004; Katzfuss and Guinness, 2017; Guinness, 2018). Des développements logiciels accompagnent ces travaux (Finley, Datta & Banerjee, 2017; Guinness, 2018).
Nous avons repris le modèle de covariance non stationnaire développé par Christopher Paciorek dans sa thèse de doctorat (Paciorek, 2003). Notre architecture de modèle est inspirée par Heinonen *et al.* (2016). Nous nous sommes aussi inspirés des modèles et applications développés par Fuglstad *et al.* (2015*a*,*b*); Ingebrigtsen, Lindgren & Steinsland (2014) et utilisant des équations différentielles partielles stochastiques, et Risser & Calder (2015) qui utilise et simplifie la covariance de Paciorek.
Les stratégies MCMC que nous développons dans l'article ont deux sources. D'une part, nos propres travaux (Coube and Liquet, 2020), qui sont eux-mêmes une application aux NNGP de l'échantillonnage chromatique de Gonzalez *et al.* (2011). D'autre part, nous avons trouvé une adaptation des méthodes hamiltoniennes de Heinonen *et al.* (2016). Ces deux types d'algorithmes sont améliorés en utilisant les entremêlements (*interweaving*) de paramétrisation de Yu and Meng (2011) et Filippone *et al.* (2013).

Bibliographie

Coube, S. & Liquet, B. 2020. Improving performances of mcmc for nearest neighbor gaussian process models with full data augmentation. *arXiv preprint arXiv:2010.00896*.

Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. 2016. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Filippone, M., Zhong, M. & Girolami, M. 2013. A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning*, 93(1):93–114.

Finley, A., Datta, A. & Banerjee, S. 2017. spnngp: spatial regression models for large datasets using nearest neighbor gaussian processes. *R package version 0.1*, 1.

Finley, A.O., Datta, A., Cook, B.D., Morton, D.C., Andersen, H.E. & Banerjee, S. 2019. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414.

Fuglstad, G.-A., Lindgren, F., Simpson, D. & Rue, H. 2015*a*. Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133.

Fuglstad, G.-A., Simpson, D., Lindgren, F. & Rue, H. 2015*b*. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531.

Gonzalez, J.E., Low, Y., Gretton, A. & Guestrin, C. 2011. Parallel gibbs sampling: From colored fields to thin junction trees. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Guinness. 2018. Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4):415–429. doi:10.1080/00401706.2018.1437476.
Available at: `https://doi.org/10.1080/00401706.2018.1437476`

Guinness, K. 2018. *GpGp: Fast Gaussian Process Computation Using Vecchia's Approximation*.
Available at: `https://CRAN.R-project.org/package=GpGp`

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S. & Lähdesmäki, H. 2016. Non-stationary gaussian process regression with hamiltonian monte carlo. *Artificial Intelligence and Statistics*.

Ingebrigtsen, R., Lindgren, F. & Steinsland, I. 2014. Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38.

Katzfuss, M. & Guinness, J. 2017. A general framework for Vecchia approximations of Gaussian processes. *arXiv e-prints*, page arXiv:1708.06302.

Lauritzen, S.L. 1996. *Graphical models*. Oxford Statistical Science Series. OUP. ISBN 9780198522195,0198522193.
Available at: `http://gen.lib.rus.ec/book/index.php?md5=7ECA79CDE5FF909E7E0E7FC8A02D8A80`

Paciorek, C.J. 2003. Nonstationary gaussian processes for regression and spatial modelling. Ph.D. thesis, Citeseer.

Risser, M.D. & Calder, C.A. 2015. Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26(4):284–297.

Stein, M.L., Chi, Z. & Welty, L.J. 2004. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Stevens, S.S. *et al.*. 1946. On the theory of scales of measurement.

Vecchia, A.V. 1988. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Yu, Y. & Meng, X.-L. 2011. To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.

PRÉVISION DANS LE MODÉLE LINÉAIRE FONCTIONNEL EN PRÉSENCE DE DONNÉES MANQUANTES DANS LA RÉPONSE ET LA COVARIABLE

Christophe Crambes[1] & Chayma Daayeb[1,2] & Ali Gannoun[1] & Yousri Henchiri[2,3]

[1]*Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.*
[2]*Université de Tunis El Manar, Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (ENIT-LAMSIN), Tunisie.*
[3]*Université de la Manouba, Institut Supérieur des Arts Multimédia de la Manouba (ISAMM), Tunisie.*
E-mail : christophe.crambes@umontpellier.fr, chayma.daayeb@etu.umontpellier.fr, ali.gannoun@umontpellier.fr, yousri.henchiri@umontpellier.fr

Résumé. Les valeurs manquantes sont un des problèmes qui surviennent fréquemment dans le processus d'observation ou d'enregistrement des données. Dans ce travail, nous considérons le modèle de régression linéaire fonctionnelle, lorsque la variable d'intérêt, réelle, et la variable explicative, fonctionnelle, contiennent des valeurs manquantes. Nous utilisons un opérateur de reconstruction qui vise à reconstruire les parties manquantes dans les courbes, puis nous nous intéressons à la méthode d'imputation par régression des données manquantes sur la variable réponse, en utilisant la régression fonctionnelle sur composantes principales pour estimer le coefficient fonctionnel du modèle. Nous étudions le comportement asymptotique de l'erreur de prévision commise lorsque les valeurs manquantes sont remplacées par les valeurs imputées. Le comportement de la méthode est également étudié en pratique sur des données simulées et réelles.

Mots-clés. Modèle linéaire fonctionnel, Données manquantes, Composantes Principales Fonctionnelles, Missing At Random, Missing Completely At Random, Imputation par régression.

Abstract. Dealing with missing values is an important issue in data observation or data recording process. In this paper, we consider a functional linear regression model, when some observations of the real response and the functional covariate are affected by missing data. We use a reconstruction operator that aims at recovering the missing parts of the explanatory curves, then we are interested in regression imputation method of missing data on the response variable, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behaviour of the prediction error we commit when missing data are replaced by the imputed values. The practical behaviour of the method is also studied on simulated and real datasets.

Keywords. Functional linear model, Missing data, Functional Principal Components, Missing At Random, Missing Completely At Random, Regression imputation.

1 Introduction

L'analyse des données fonctionnelles a connu un développement très important ces dernières années, comme en attestent les nombreux ouvrages sur le sujet : Ramsay et Silverman (2005), Ferraty et Vieu (2006), Hsing et Eubank (2015) constitue une liste non exhaustive de monographies donnant une vision d'ensemble sur ce thème. Un des modèles les plus populaires en analyse de données fonctionnelles est le modèle linéaire fonctionnel, qui établit une relation de dépendance entre une variable réelle Y et une variable aléatoire fonctionnelle $X = (X(t), t \in [a, b])$. La variable X est à valeurs dans l'espace $H := L^2([a, b])$ des fonctions de carré intégrable sur l'intervalle compact $[a, b]$. Nous supposons dans la suite que $\mathbb{E}(\|X\|^2) < +\infty$ où $\|.\|$ désigne la norme usuelle de H associée au produit scalaire $\langle ., . \rangle$, défini par $\langle f, g \rangle = \int_a^b f(t)g(t)\mathrm{d}t$ pour toutes fonctions f et g de H. Le modèle linéaire fonctionnel a été étudié par de nombreux auteurs, par exemple Cardot et al. (1999), Cai et Hall (2006), Hall et Horowitz (2007), Crambes et al. (2009). Ce modèle est défini par

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon, \tag{1.1}$$

où $\theta_0 \in \mathbb{R}$ et $\theta \in H$ sont les paramètres à estimer. L'erreur du modèle ε est une variable aléatoire réelle centrée indépendante de X avec une variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. Le modèle (1.1) peut s'écrire sous la forme

$$Y = \theta_0 + \Theta X + \varepsilon, \tag{1.2}$$

où $\Theta : H \to \mathbb{R}$ est l'opérateur linéaire continu défini par $\Theta u = \langle \theta, u \rangle$ pour toute fonction $u \in H$. Dans la suite, nous considérons un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ indépendant et identiquement distribué de même loi que le couple (X, Y). Pour estimer θ ou Θ, nous considérons la régression fonctionnelle sur composantes principales. Il s'agit d'une régression des moindres carrés de la réponse Y sur les variables réelles qui sont les coordonnées de la projection de X sur l'espace engendré par les fonctions propres associées aux plus grandes valeurs propres de l'opérateur de covariance de X (voir Cardot et al., 1999). Soit $(k_n)_{n\geq 1}$ une suite de nombres entiers, l'estimateur $\widehat{\Theta}$ de Θ proposé par Cardot et al. (1999) est défini par

$$\widehat{\Theta} = \langle \widehat{\theta}, . \rangle = \widehat{\Pi}_{k_n} \widehat{\Delta}_n (\widehat{\Pi}_{k_n} \widehat{\Gamma}_n \widehat{\Pi}_{k_n})^{-1}, \tag{1.3}$$

où $\widehat{\Delta}_n$ est l'opérateur de covariance croisée empirique donné par $\widehat{\Delta}_n u = \frac{1}{n} \sum_{i=1}^n \langle X_i, u \rangle Y_i$ pour tout $u \in H$, $\widehat{\Gamma}_n$ est l'opérateur de covariance empirique défini par $\widehat{\Gamma}_n u = \frac{1}{n} \sum_{i=1}^n \langle X_i, u \rangle X_i$ pour tout $u \in H$ et $\widehat{\Pi}_{k_n} = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u \rangle \widehat{\phi}_j$ est l'opérateur de projection orthogonale sur le sous-espace engendré par les fonctions propres $(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n})$ associées aux k_n plus grandes valeurs propres $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{k_n}$ de l'opérateur $\widehat{\Gamma}_n$. En supposant que $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ p.s., l'estimateur de θ est donné par

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{\phi}_j\rangle Y_i}{\widehat{\lambda}_j}\widehat{\phi}_j. \tag{1.4}$$

En outre, l'estimateur de $\theta_0 = \mathbb{E}(Y) - \int_a^b \theta(t)\mathbb{E}(X(t))\mathrm{d}t$ s'écrit sous la forme $\widehat{\theta}_0 = \overline{Y} - \int_a^b \widehat{\theta}(t)\overline{X}(t)\mathrm{d}t$ où $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ et $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Le cadre de ce travail est la situation où des valeurs manquantes affectent à la fois la réponse et la variable explicative. L'objectif est ***(i)*** de reconstituer les courbes X manquantes et d'imputer les données manquantes sur Y, ***(ii)*** d'estimer θ ou Θ avec le jeu de données reconstitué, ***(iii)*** de prédire une nouvelle valeur de la réponse Y étant donnée une nouvelle observation test sur la variable explicative X.

2 Données manquantes

Pour le mécanisme de données manquantes dans la réponse, nous considérons une variable aléatoire binaire $\delta^{[Y]}$ et un échantillon $(\delta_i^{[Y]})_{i=1,\ldots,n}$ tel que $\delta_i^{[Y]} = 1$ si la valeur Y_i est observée et $\delta_i^{[Y]} = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \ldots, n$. Nous considérons les données manquantes de la réponse "Missing At Random" (MAR) : le fait que la valeur Y est manquante ne dépend pas de la réponse du modèle, mais peut éventuellement dépendre de la covariable, c'est-à-dire

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

Dans ce qui suit, le nombre de valeurs manquantes parmi $Y_1, \ldots, Y_n$ est noté

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\left\{\delta_i^{[Y]}=0\right\}}.$$

Dans Crambes et Henchiri (2019), une méthodologie d'imputation des données manquantes par régression est donnée, sous cette hypothèse MAR, mais la covariable est censée être complètement observée, ce qui n'est plus le cas ici. Nous considérons une variable fonctionnelle $\delta^{[X]}$ et un échantillon $(\delta_i^{[X]})_{i=1,\ldots,n}$ tel que, pour $t \in [a, b]$, $\delta_i^{[X]}(t) = 1$ si $X_i(t)$ est observé et $\delta_i^{[X]}(t) = 0$ si $X_i(t)$ est manquant. Nous considérons les données manquantes de la covariable "Missing Completely At Random" (MCAR) : le fait que X contient des données manquantes ne dépend pas de la covariable du modèle, ni de la réponse, c'est-à-dire, pout tout $t \in [a, b]$

$$\mathbb{P}(\delta^{[X]}(t) = 1 \mid X, Y) = \mathbb{P}(\delta^{[X]}(t) = 1).$$

D'autre part, le nombre de courbes où des valeurs manquantes apparaissent est donné par

$$m_n^{[X]} = \sum_{i=1}^{n} \mathbb{1}_{\left\{\exists\, t\in\,[a,b],\, \delta_i^{[X]}(t)=0\right\}}.$$

Dans la suite, nous présentons la méthodologie de reconstruction des courbes, puis l'imputation par régression d'une valeur manquante pour la variable d'intérêt. Enfin, nous donnons l'estimateur de θ à partir du jeu de données reconstitué, et la prédiction d'une valeur de la réponse suite à la donnée d'une nouvelle observation test pour la variable explicative.

3 Reconstruction des covariables manquantes

Nous appliquons dans cette partie la méthodologie introduite dans Kneip et Liebl (2020) pour reconstruire la partie manquante d'une courbe. Soit $(O_i)_{i=1,\dots,n}$ l'échantillon des périodes d'observation des courbes, c'est-à-dire $O_i = \left\{ t \in [a,b], \delta_i^{[X]}(t) = 1 \right\}$ pour tout $i = 1, \dots, n$. En outre, notons $M_i = [a,b] \setminus O_i$ pour tout $i = 1, \dots, n$. Dans la suite, nous utilisons O et M pour désigner une production donnée de O_i et M_i. De plus, nous notons la partie observée de X_i par X_i^O et X_i^M pour la partie manquante. Nous considérons la décomposition de Karhunen-Loève (KL) pour X_i^O dans $L^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t), \tag{3.1}$$

pour $t \in O$, où $(\phi_k^O)_{k\geq 1}$ désigne la suite des fonctions propres de l'opérateur de covariance de X_i^O et $(\xi_{ik}^O)_{k\geq 1}$ est une suite de variables aléatoires centrées et décorrélées avec $\mathbb{E}(\xi_{ik}^O) = \lambda_k^O$, la suite $(\lambda_k^O)_{k\geq 1}$ étant la suite des valeurs propres de l'opérateur de covariance de X_i^O. La partie manquante de la courbe s'écrit, pour $t \in O$ et $s \in M$

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s), \tag{3.2}$$

où $L : L^2(O) \to L^2(M)$ est un opérateur linéaire défini par

$$L(X_i^O(t)) = \sum_{k=1}^{+\infty} \frac{\mathbb{E}[\xi_{ik}^O X_i^M(s)]}{\lambda_k^O},$$

dont le but est de reconstruire les parties manquantes $X_i^M \in L^2(M)$ à partir des observations $X_i^O \in L^2(O)$. La variable $Z_i \in L^2(M)$ est l'erreur de reconstruction. Nous cherchons à minimiser l'erreur quadratique moyenne $\mathbb{E}[(X_i^M(t) - L(X_i^O)(t))^2]$ avec $t \in M$, pour obtenir l'opérateur de reconstruction linéaire optimal suivant les étapes : (**I**) estimation par polynômes locaux des courbes X et de la fonction de covariance, (**II**) estimation des valeurs propres et des fonctions propres de l'opérateur de covariance de X sur la partie observée des courbes, (**III**) estimation sur la partie manquante de la courbe à l'aide des éléments propres de la partie observée. Sous des hypothèses classiques dans ce contexte, des vitesses de convergence uniforme de la courbe reconstruite vers la vraie courbe sont données dans Kneip et Liebl (2020).

Dans la suite, nous notons $\widehat{X}$ une courbe X reconstruite et $X^\star = \delta^{[X]} X + (1 - \delta^{[X]})\widehat{X}$.

4 Imputation par régression

Nous nous intéressons ici à l'imputation des données manquantes sur la réponse Y, suivant la méthode présentée dans Crambes et Henchiri (2019). Nous définissons l'opérateur de covariance avec les courbes reconstruites par

$$\widehat{\Gamma}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} X_i^\star.$$

Soit ℓ un nombre entier compris entre 1 et n tel que Y_ℓ soit manquante, c'est-à-dire avec $\delta_\ell^{[Y]} = 0$. La valeur imputée par régression pour Y_ℓ est définie par

$$Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle, \tag{4.1}$$

avec

$$\widetilde{\theta} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i}{\widehat{\lambda}_{j,rec}^{obs}} \widehat{\phi}_{j,rec}^{obs} \quad \text{et} \quad \widetilde{\theta}_0 = \overline{Y}_{obs} - \int_a^b \widetilde{\theta}(t) \overline{X}^\star(t) \mathrm{d}t,$$

où $(\widehat{\lambda}_{j,rec}^{obs})_{j \geq 1}$ et $(\widehat{\phi}_{j,rec}^{obs})_{j \geq 1}$ sont les éléments propres de l'opérateur $\widehat{\Gamma}_{n,rec}^{obs}$ et les moyennes empiriques sont $\overline{Y}_{obs} = \frac{1}{n} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i$ et $\overline{X}^\star = \frac{1}{n} \sum_{i=1}^{n} X_i^\star$.

Sous des hypothèses analogues à celles de Kneip et Liebl (2020) et Crambes et Henchiri (2019), nous obtenons des vitesses de convergence pour la valeur imputée $Y_{\ell,imp}$ de façon similaire à Crambes et Henchiri (2019) (cas d'une covariable fonctionnelle complètement observée et d'une réponse affectée par des données manquantes).

5 Prédiction

Une fois la base de données reconstruite, nous estimons le coefficient fonctionnel θ et l'intercept θ_0, respectivement, par

$$\widehat{\theta}^\star = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec} \rangle Y_i^\star}{\widehat{\lambda}_{j,rec}} \widehat{\phi}_{j,rec} \quad \text{et} \quad \widehat{\theta}_0^\star = \overline{Y}^\star - \int_a^b \widehat{\theta}^\star(t) \overline{X}^\star(t) \mathrm{d}t,$$

avec $Y_i^\star = Y_i \delta_i^{[Y]} + Y_{i,imp}(1 - \delta_i^{[Y]})$, pour tout $i = 1, \ldots, n$, et $\overline{Y}^\star = \frac{1}{n} \sum_{i=1}^{n} Y_i^\star$. Nous pouvons à présent définir la prédiction d'une nouvelle valeur Y_{new} associée à l'observation X_{new} de la variable explicative par

$$\widehat{Y}_{new} = \widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle. \tag{5.1}$$

Une étude asymptotique de l'erreur de prévision de $\widehat{Y}_{new}$ a été réalisée. Le comportement de la méthode en pratique a également été évalué sur des données simulées et réelles. À titre d'exemple, nous avons simulé 400 réplications sur le modèle (1.1) avec $[a;b] = [0;1]$, $\theta_0 = 3$, et $\theta(t) = \sum_{j=1}^{50} b_j \Phi_j(t)$ pour tout $t \in [0;1]$, où $b_1 = 0.3$, $b_j = 4(-1)^{j+1} j^{-2}$ pour $j > 1$ et $\Phi_1(t) = 1$, $\Phi_j(t) = \sqrt{2}\cos(j\pi t)$ pour $t \in [0;1]$ et $j > 1$. Le bruit ε est simulé suivant la loi $N(0;\sigma_\varepsilon^2)$ avec $\sigma_\varepsilon^2 = 0.04$ et la variable X est simulée en écrivant $X(t) = \sum_{j=1}^{150} \xi_j \lambda_j \Phi_j(t)$ où $\lambda_j = (-1)^{j+1} j^{-2}$ pour $j \geq 1$ et ξ_j suit la loi uniforme sur l'intervalle $[-\sqrt{3};\sqrt{3}]$. Nous avons pris une taille d'échantillon $n = 360$ et les courbes ont été discrétisées à l'aide de $p = 100$ points de mesure. Concernant les données manquantes, les parties $[0;1/8]$ et $[7/8;1]$ ont été retirées aléatoirement de 14.8% des courbes X et 12% de données ont été retirées de Y. Sur les 400 réplications, nous avons calculé une erreur quadratique moyenne de prévision de 8.46×10^{-2} avec un écart-type de 8.66×10^{-2} lorsque nous avons reconstruit les courbes X et imputé les données manquantes sur Y. L'erreur quadratique moyenne de prévision devient 9.24×10^{-2} avec un écart-type de 8.82×10^{-2} lorsque l'on se contente de supprimer les observations pour lesquelles X ou Y est manquant. Cela montre, sur cet exemple l'intérêt de reconstituer le jeu de données plutôt que de simplement enlever les données manquantes.

Bibliographie

CAI, T.T. and HALL, P. (2006). Prediction in functional linear regression. *Annals of Statistics*, **34**, 2159-2179.

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statistics and Probability Letters*, **45**, 11-22.

CRAMBES, C. and HENCHIRI, Y. (2019). Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, **201**, 103-119.

CRAMBES, C. KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of statistics*, **37**, 35-72.

FERRATY, F. and VIEU, P. (2006). *Nonparametric functional data analysis : Theory and practice*. Springer-Verlag, New York.

HALL, P. and HOROWITZ, J.L. (2007). Methodology and convergence rates for functional linear regression, *The Annals of Statistics*, **35**, 70-91.

HSING, T. and EUBANK, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics, John Wiley & Sons, Chichester.

KNEIP, A and LIEBL, D. (2020). On the optimal reconstruction of partially observed functional data. *Annals of Statistics*, **4**, 1692-1717.

RAMSAY, J.O. and SILVERMAN, B.W. (2005). *Functional Data Analysis* (Second edition). Springer-Verlag, New York.

Using Random forest and Gradient boosting trees to improve wave forecast at a specific location

Aurélien Callens [1], Denis Morichon [2], Stéphane Abadie [2], Matthias Delpey [3] & Benoit Liquet [1]

[1] *Université de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France; aurelien.callens@univ-pau.fr / benoit.liquet@univ-pau.fr ; 1 Allée du Parc Montaury, 64600 Anglet.*

[2] *Université de Pau et des Pays de l'Adour, E2S UPPA, SIAME, Anglet, France; denis.morichon@univ-pau.fr / stephane.abadie@univ-pau.fr ; 1 Allée du Parc Montaury, 64600 Anglet.*

[3] *Centre Rivages Pro Tech SUEZ EAU FRANCE, Bidart, France; matthias.delpey@suez.com ; 2 Allée Théodore Monod F, Izarbel 64210 Bidart.*

Résumé.

Les modèles de vagues déterministes comme WW3 ou MFWAM sont couramment utilisés pour prédire les caractéristiques des vagues au niveau de la côte. Il a été montré que durant des conditions climatiques extrêmes (vents forts et tempêtes), ces modèles ont tendance à sous-estimer les valeurs de certains paramètres de vagues notamment la hauteur significative des vagues. La solution actuelle pour ce problème est d'utiliser la méthode de prédiction des erreurs qui est une méthode d'assimilation de données se basant sur des modèles d'apprentissage statistique et visant à réduire les erreurs de prédiction. Dans le domaine l'ingénierie côtière, la méthode des réseaux de neurones est la méthode par défaut pour la technique de prédiction des erreurs. L'objectif principal de ce travail est de montrer deux alternatives aux réseaux de neurones : les forêts aléatoires et les arbres de décision boostés. Nous avons montré que le RMSE des paramètres de vagues améliorés par les forêts aléatoires et arbres de décision boostés sont respectivement 10 et 20% plus petits que les RMSE obtenus avec les réseaux de neurones, indiquant ainsi de meilleures performances (Callens et al., 2020). L'objectif secondaire de ce travail est de montrer comment optimiser au mieux les hyperparamètres des algorithmes de machine learning avec la méthode d'Optimisation Bayésienne. L'optimisation des hyperparamètres est une étape essentielle lors de l'utilisation d'une méthode d'apprentissage statistique car elle améliore généralement les résultats de manière significative. C'est le cas pour notre étude, où l'optimisation des hyperparamètres a permis d'avoir des valeurs RMSE plus faibles, en moyenne entre 8 et 11% pour la correction de la hauteur significative des vagues et de leur période.

Mots-clés. Arbres de décision boostés, Assimilation de données, Forêt aléatoires, Prédiction d'erreurs, Prévision des paramètres de vague, Réseaux de neurones artificiels.

Abstract.

In coastal engineering, wind wave models such as WW3 or MFWAM are commonly used to predict the wave characteristics at the shore. It has been proven that during extreme conditions (high winds and storm), they have a tendency to underestimate certain wave parameters especially significant wave height. The current solution for this problem is to improve the wave parameter forecast with the the error prediction method which is a data assimilation technique based on statistical learning models and aiming to reduce prediction errors. In the field of coastal engineering, the method of neural networks is the off-the-shelf method for error prediction method. The main objective of this work is to present two alternative algorithms to neural networks, namely random forest and gradient boosting tree. We showed that the RMSE of the variables updated with gradient boosting trees and random forest are respectively 20 and 10% lower than the RMSE obtained with neural networks (Callens et al., 2020). A secondary objective is to show how to tune the hyperparameter values of machine learning algorithms with Bayesian Optimization. This step is essential when using machine learning algorithms and can improve the results significantly. Indeed, after a fine hyperparameter tuning with Bayesian optimization, gradient boosting trees yielded RMSE values in average 8% to 11% lower for the correction of significant wave height and peak wave period.

Keywords. Artificial neural networks, Data assimilation, Error prediction, Gradient boosting trees, Random forest, Wave forecasting.

1 Introduction

Nowadays, numerical wave models are routinely used to forecast wind generated waves. Although they provide satisfactory predictions at a regional scale and during mean wave conditions, it has been shown that they are less accurate for forecasting at a specific location (Londhe et al., 2016) and have a tendency to underestimate wave height during energetic wave conditions.

When observation data are available, data assimilation can be used to improve the predictions made by numerical models. There are 4 main categories of data assimilation procedures (Babovic et al., 2001): updating the input parameters, updating the state variables, updating the model parameters and finally updating the output parameters. The last procedure is called "Error prediction" method and is the most suitable approach to improve model predictions of different output variables at a specific location (Babovic et al., 2005).

This method has been successfully applied on hindcast data (Makarynskyy et al., 2005) and has even been implemented in real time setting in the works of Babovic et al. (2001) and Londhe et al. (2016). To our knowledge, only artificial neural networks have been tested to forecast the errors in the data assimilation. However, according to the so-called "No Free Lunch" theorem, there is no single model that works best for all problems (Wolpert, 2002). It is therefore necessary to try multiple models and find the one that

works best for our particular problem.

Random forest and gradient boosting trees are strong candidates for comparison with neural networks. Indeed, these two methods are known for their performance and unlike neural networks, they also provide valuable information by computing the predictive power of each variable used as input. This study aims to present two alternatives (random forest and gradient boosting trees) to neural networks by comparing their performances when improving regional numerical models. A secondary objective is to show how to tune the hyperparameter values of machine learning algorithms with Bayesian Optimization.

2 Data and Method

2.1 Data

- Measured wave parameters (H_s: significant wave height, T_P: Peak wave period, θ_P: peak wave direction) : obtained from a buoy located a few miles off Biarritz Coast. From 2009 to nowadays, 30 min time-step.
- Modelled wave parameters (H_s, T_p, θ_p) with MFWAM model (Lefèvre and Aouf, 2012). These data were chosen due to their free accessibility on Copernicus website. From 2007 to 2019, hourly time-step.
- Meteorological conditions (Wind speed, direction and atmospheric pressure): MeteoFrance weather station in Biarritz. From 2013 to 2018, hourly time step.

By assembling the wave buoy data, the wind wave parameters and the meteorological data we obtain a dataset of 41439 hourly observations ranging from 2013-01-01 to 2018-12-31. In this work, we are improving the wave forecast by correcting the systematic errors of the wind wave model. Therefore, we are not considering any temporal effects while improving H_S, T_p and θ_p. The assimilation made in this work is only valid for the buoy located a few miles off Biarritz, it can be extended to other buoys by integrating their data to this analysis. The dataset was randomly divided into 2 parts: the training part containing 70% of the observations ($n = 28797$) and the testing part containing the remaining 30% ($n = 12342$).

2.2 Method

The error prediction method consists in three steps:

- Step 1: Deviations between model predictions and measured values are computed:

$$E_{model} = X_{measured} - X_{modeled},$$

where E_{model} is the error of the model, $X_{measured}$ is the measured value of an output variable provided by the wave buoy and $X_{modeled}$ is the value of the same variable computed by the wave model.

- Step 2: E_{model} is predicted with an appropriate supervised machine learning algorithm.

- Step 3: The predicted error is added to the prediction of the wave model to obtain an updated numerical prediction:

$$X_{updated} = X_{modeled} + E_{predicted},$$

where $X_{updated}$ is the updated prediction of wave model and $E_{predicted}$ is the predicted error given by the supervised learning method.

This method is repeated separately for each output variable to improve (H_s, T_p, θ_p) which is usual in the literature (Moeini et al., 2012). Concerning the statistical learning method, only neural networks have been used for the step (2) of the error prediction method to our knowledge(Moeini et al., 2012; Londhe et al., 2016). Because we want to compare the performance between different machine learning algorithms, we use random forest and gradient boosting trees. All the tested algorithms use the same input variables to improve the model accuracy: the three wave parameters (H_S, T_p, θ_p) given by the numerical model, the atmospheric pressure, the wind direction and speed.

The same learning algorithm is tested twice in the error prediction method: once with the default hyperparameter values and once with the optimal hyperparameter values chosen by Bayesian optimisation. Full details on Bayesian optimisation are given in the work of Snoek et al. (2012).

3 Results

The results obtained for the H_s parameter on the test set are shown in table 1. The numerical model shows a negative bias, indicating that the MFWAM model has a tendency to underestimate H_s such as other wind wave models (Moeini et al., 2012). This negative bias increases as the value of H_s becomes larger ($H_s > 3m$), meaning that H_s is more likely to be underestimated during energetic events. With the assimilation technique, the different algorithms remove the bias observed for H_s (all data). For the data where $H_s > 3m$, the correction does not remove the bias for H_s but reduces it greatly. The best algorithm with default values of hyperparameters is random forest followed closely by gradient boosting trees. After tuning the hyperparameter, the values of bias and RMSE are slightly reduced for all algorithms and gradient boosting trees becomes the best method for the improvement of H_s.

Table 1: Statistical metrics for the three variables of interest before the hyperparameter tuning. "Ann" stands for artificial neural networks, "Rf" for random forest and "Gb" for gradient boosting tree.

	Hs (Default values)				Hs (Bayesian optimisation)			
	Numerical model	Ann Corr.	Rf Corr.	Gb Corr.	Numerical model	Ann Corr.	Rf Corr.	Gb Corr.
	Computed with all data							
Biais	-0.201	0.005	-0.002	-0.004	-0.201	0.026	-0.002	-0.001
RMSE	0.399	0.306	0.248	0.267	0.399	0.300	0.246	0.240
	Computed with data where $H_s > 3m$							
Biais	-0.536	-0.156	-0.124	-0.126	-0.536	-0.117	-0.120	-0.099
RMSE	0.766	0.515	0.420	0.433	0.766	0.495	0.417	0.404

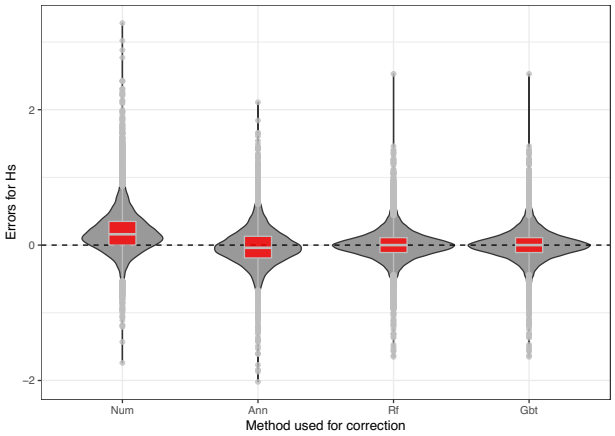


Figure 1: Distribution of the H_s errors computed between values observed at the buoy and values corrected or not with the different machine learning algorithms. "Num" stands for numerical model (no correction), "Ann" for artificial neural networks, "Rf" for random forest and "Gb" for gradient boosting trees.

The distribution of the errors for H_s parameter after the different corrections are presented in the figure 1. The distributions of the errors after a correction have narrowed and are now more centered in zero. The differences in performance between algorithms are confirmed with these violin plots. Indeed, when the correction is made with random forest or gradient boosting trees, the distributions of the errors are narrower than the distributions of the errors obtained with neural networks.

4 Conclusion

Random forest and Gradient boosting trees performed better than neural network in the error prediction method. Indeed, the best algorithm for our data is gradient boosting trees with RMSE reduced by 40% for H_s, 33% for T_p and 31% for θ_p. Concerning hyperparameter tuning, bayesian optimisation led to better results: it lowered the RMSE values obtained by gradient boosting trees by 8 to 11% in average for the three wave parameters (H_s, T_p, θ_p).

Bibliography

Babovic, V., Caňizares, R., Jensen, H. R., and Klinting, A. (2001). Neural networks as routine for error updating of numerical models. *Journal of Hydraulic Engineering*, 127(3):181–193.

Babovic, V., Sannasiraj, S. A., and Chan, E. S. (2005). Error correction of a predictive ocean wave model using local model approximation. *Journal of Marine Systems*, 53(1-4):1–17.

Callens, A., Morichon, D., Abadie, S., Delpey, M., and Liquet, B. (2020). Using random forest and gradient boosting trees to improve wave forecast at a specific location. *Applied Ocean Research*, 104:102339.

Lefèvre, J.-M. and Aouf, L. (2012). Latest developments in wave data assimilation. In *ECMWF Workshop on Ocean Waves*, pages 25–27.

Londhe, S. N., Shah, S., Dixit, P. R., Nair, T. M. B., Sirisha, P., and Jain, R. (2016). A Coupled Numerical and Artificial Neural Network Model for Improving Location Specific Wave Forecast. *Applied Ocean Research*, 59:483–491.

Makarynskyy, O., Pires-Silva, A. A., Makarynska, D., and Ventura-Soares, C. (2005). Artificial neural networks in wave predictions at the west coast of Portugal. *Computers & Geosciences*, 31(4):415–424.

Moeini, M. H., Etemad-Shahidi, A., Chegini, V., and Rahmani, I. (2012). Wave data assimilation using a hybrid approach in the Persian Gulf. *Ocean Dynamics*, 62(5):785–797.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959.

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft Computing and Industry*, pages 25–42. Springer.

Probabilistic expert aggregation via additive stacking

Christian Capezza [1], Biagio Palumbo [2], Yannig Goude [3], Simon N. Wood [4] & Matteo Fasiolo [5],

[1] *Department of Industrial Engineering, University of Naples Federico II, Naples, Italy, christian.capezza@unina.it*
[2] *Department of Industrial Engineering, University of Naples Federico II, Naples, Italy, biagio.palumbo@unina.it*
[3] *Electricité de France, Clamart, France, yannig.goude@edf.fr*
[4] *School of Mathematics, University of Edinburgh, Edinburgh, UK, simon.wood@ed.ac.uk*
[5] *School of Mathematics, University of Bristol, Bristol, UK, matteo.fasiolo@bristol.ac.uk*

Abstract. It is particularly challenging to forecast the individual household eletricity demand because of its lower signal-to-noise ratio compared to the aggregate demand and the increased variability due to the increasing use of renewable energy sources and distributed production. However, accurate forecasts will be essential for a profitable smart grid management. We propose a new method for probabilistic expert aggregation, which borrows information between the different households, taking their individual characteristics into account. The main innovation of the propose method is that the experts' weights vary depending on covariates through an additive model structure.

Keywords. Generalized additive models, probabilistic forecasting, ensemble forecasting, accumulated local effect plots

1 Additive Stacking

In this work we propose additive stacking for probabilistic forecasting of electricity demand at household level. Existing methods for short term load forecasting of the aggregate demand have reached high levels of accuracy, but cannot be used directly to forecast disaggregate demand, which is characterized by a much lower signal-to-noise ratio and abrupt changes in the demand distribution. Probabilistic additive stacking allows us to model the full demand distribution, which is more practically valuable than simple point estimates when dealing with very noisy household level data. We propose a stacking of experts that combines several probabilistic forecasts with heterogeneous characteristics, with the aim of providing improved probabilistic forecasting accuracy. Yao et al. (2018) proposed stacking to average Bayesian predictive distributions, however weights of the stacking are fixed. As first innovation of our work, we allow the weights of the combination of experts to vary depending on covariates through an additive model, where effects

can be smooth functions. The estimation of coefficients and smoothing parameters is conveniently performed using Wood, Pya and Säfken (2016), which proposed a general framework for smoothing parameter estimation for models with regular likelihoods constructed in terms of unknown smooth functions of covariates. In order to demonstrate the performance of the proposed method, we used the Irish smart meter data set from the Commission for Energy Regulation (CER) Smart Metering Project (Commission for Energy Regulation (2012)), which provides one year's worth of data on electricity demand at individual household level. Given that we are dealing with a large data set, comprising demand at 30 minutes resolution for several thousand customers, we exploit the Big Data tool provided by the mgcv R package to fit the proposed additive stacking model to the whole data set.

The second innovation of this work is the use of new visualization methods for interpreting the effect of covariates on the weights of stacking. In fact, since weights of stacking depend on covariates through opportune transformations of linear predictors to preserve non negativity and sum-to-one constraints, by simply estimating the stacking model it is not immediate to visualize the effect of covariates on the weights directly, then smooth effects alone are not interpretable. We rely on accumulated local effect (ALE) plots introduced by Apley (2016), which however does not provide information about their uncertainty. Therefore, we propose an approximate method to obtain credible intervals of accumulated local effects. As an example of ALE plots, in Figure 1 we show an implementation on the CER data set, where we perform an aggregation of four experts: $LastMonth$, which uses as probabilistic forecast a kernel density estimate based on the observations in the previous month; $GaulssInd$, a generalized additive model for location, scale and shape (GAMLSS), estimated separately on each household; $Dynamic$, an adaptive generalized additive model, estimated using only the last three days data of each household; $GaulssCommon$, a GAMLSS using data from all customers. Two simple conclusions can be drawn in Figure 1, by looking at the effect of two covariates on the aggregation weights: the effect of the time of the day gives largest weight to $GaulssInd$, which is the most complex model, in the day time and to $LastMonth$ in the night; the $Dynamic$ expert, which adapts quickly using only the most recent data, gets a large weight when a customer was out of home in the previous days.

In the proposed application we evaluate the predictive performance of the proposed method, relative to the experts, showing that additive stacking outperforms all experts under several loss functions (e.g., the pinball loss, mean square error, continuous ranked probability score). However, these standard loss functions for evaluating the forecasting performance at disaggregate level do not really convey or quantify the practical utility of producing an accurate disaggregated probabilistic forecast. Therefore, we evaluate the performance of the proposed approach also using a simple application, which is meant to highlight the utility of by-customer probabilistic forecasting in the context of disaggregate demand management.

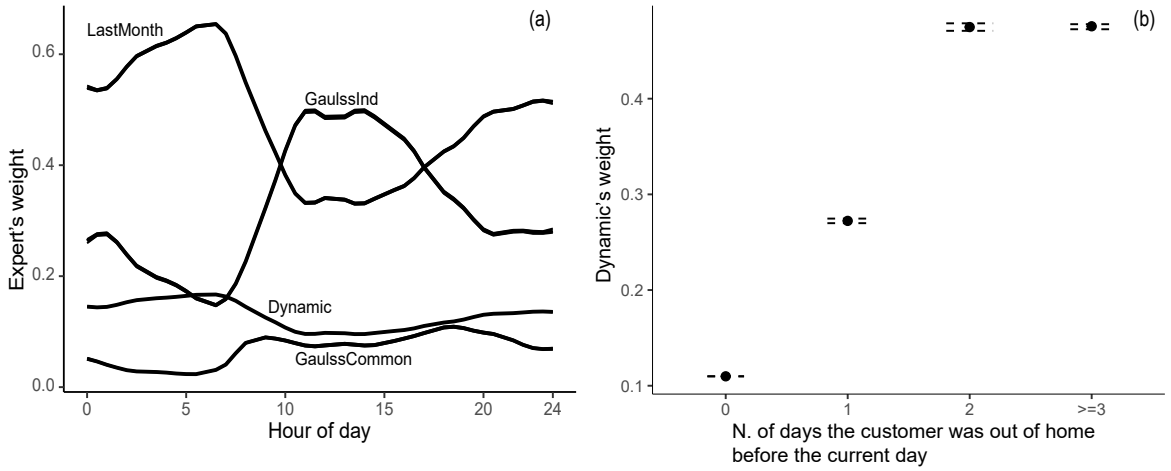


Figure 1: ALE effects on the stacking weights of some covariates: a) effect of time of day on each expert, b) effect of the number of previous days a customer was out of home on *Dynamic*.

Bibliographie

Apley, D.W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468.

Commission for Energy Regulation (CER) (2012). CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. www.ucd.ie/issda/CER-electricity

Wood, S.N., Pya, N. and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models, *Journal of the American Statistical Association*, 111(516), pp. 1548-1563.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion), *Bayesian Analysis*, 13(3), pp. 917-1007.

Sparse inverse time correlation model for signal identification in functional Near Infrared Spectroscopy data

David Causeur [1] & Ching-Fan Sheu [2]

[1] *Agrocampus Ouest, Irmar, UMR 6625 CNRS, 65 rue de St-Brieuc - CS 84215, 35042 Rennes Cedex, France, david.causeur@agrocampus-ouest.fr*

[2] *National Cheng-Kung University, Institute of Education, 1 University Road, Tainan 701, Taiwan, csheu@mail.ncku.edu.tw*

Résumé. La spectroscopie proche infra-rouge fonctionnelle (fNIRS) utilise les propriétés d'absorption par l'hémoglobine de la lumière dans la gamme du proche infra-rouge pour suivre les variations dans le temps de l'oxygénation du sang. Elle est en particulier utilisée pour mesurer l'activité cérébrale fonctionnelle de sujets au cours d'exercices conçus pour stimuler une réponse cérébrale d'intérêt au regard d'une pathologie ou d'un trouble neurologique ou psychiatrique. Ainsi, la technologie fNIRS génère des courbes décrivant la réponse hémodynamique du cerveau en temps réel, ce qui permet d'en déduire des motifs d'association avec les valeurs contrôlées de facteurs expérimentaux. Dans ce contexte, l'analyse de variance fonctionnelle (fANOVA) offre toute la flexibilité du modèle linéaire multivarié pour comparer des modèles d'association aux facteurs expérimentaux, sous des hypothèses additionnelles de régularité des motifs d'association et de dépendance temporelle entre les résidus. Causeur *et al.* (2020) démontre que la manière dont la dépendance temporelle est prise en compte dans les procédures d'analyse de la variance fonctionnelle a un impact considérable sur la puissance des tests et que la prise en compte optimale dépend de la conjonction entre la forme des motifs d'association et la répartition des corrélations temporelles. Pour la problématique d'identification d'intervalles de temps sur lesquels le signal d'association est non-nul, nous proposons une procédure d'estimation doublement pénalisée, offrant un large espace de recherche de la procédure optimale, selon la parcimonie du motif d'association et celle de la matrice de corrélation inverse des résidus.

Mots-clés. Données fonctionnelles, Estimation pénalisée, Grande dimension, Modèle de corrélation inverse, Spectroscopie Proche Infra-Rouge fonctionnelle.

Abstract. Functional near infrared spectroscopy (fNIRS) uses the absorption of near infrared light by hemoglobin to record changes in blood oxygenation as signals of functional brain activity. For designs in which subjects are instructed to execute a specific mental task under different experimental conditions with pre-determined levels for covariates, fNIRS provides real-time cerebral hemodynamic responses for studying neural correlates of task-related experimental variables. Data obtained from such designs are discretized observations of the hemodynamic curves on a high-resolution time scale. Testing for overall group mean differences among curves or, more generally, relationships between

curves and explanatory variables can be addressed by using functional Analysis of Variance (fANOVA) procedures in a general multivariate linear regression framework where additional assumptions are made to account for the regularity of mean curves and for the strong time-dependence across residuals. Causeur et al. (2020) demonstrated that how way time dependence is modeled in such fANOVA testing procedures is crucial and should account for the interplay between the pattern of regression parameter curves and the distribution of the time correlations. To address the challenging issue of identifying time points for which the association signal is nonzero, we propose a doubly penalized estimation procedure assuming that both the association signal and the inverse time correlation matrix are sparse. We show how the tuning of penalty parameters enables a flexible handling of dependence and deduce optimal signal identification procedures.

Keywords. Functional data, fNIRS data, Inverse correlation model, High-dimensional data, Penalized estimation.

Introduction

Functional data are discrete observations generated from underlying continuous functions. Progress in instrumentations and techniques for measurements has produced massive data that are functional in nature, such as measurements over fine time or space grids. Examples are changes of brain function associated with oxygen supply in blood flow in functional magnetic resonance imaging (fMRI) or functional Near Infrared Spectroscopy (fNIRS).

For a typical fNIRS study, brain hemodynamic response curves are measured in milliseconds (ms) for up to several seconds with reference to the onset of an event; namely, subject's response is a high resolution curve describing her functional brain activity in time. As in traditional statistical analysis in which the response provides a stable description of the subject through one or, at most, a few variables, functional data are similarly analyzed for covariate effects. Functional Analysis of Variance (fANOVA) is especially designed to extend standard analysis of variance (ANOVA) to situations where the response is a curve (see Causeur *et al.*, 2020, for a review).

When testing for the effect of fixed-time covariates on curves, where a random curve $t \mapsto Y(t)$ is observed on a grid of $m \geq 1$ time points t_j, $j = 1, \ldots, m$ and $\boldsymbol{Y} = (Y(t_1), \ldots, Y(t_m))$, most functional Analysis of Variance (ANOVA) methods rely on the multivariate linear regression model, with additional assumptions on the regularity over time of effect curves:

$$\boldsymbol{Y} \;=\; \boldsymbol{\beta}_0 + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)'$ is a p–vector of explanatory variables, with $p \geq 1$, $\boldsymbol{\beta}_0$ is the m–vector of intercept parameters, $\boldsymbol{\beta}$ is the $p \times m$ matrix of regression parameters and $\boldsymbol{\varepsilon}$

is the m–vector of residual error terms, assumed to be normally distributed, with mean 0 and positive variance-covariance matrix $\boldsymbol{\Sigma}$.

In most fNIRS designs, m exceeds by far the number n of independent joint observations of the explanatory and response variables but the rank p of the sample variance-covariance matrix $\boldsymbol{S}_{xx}$ of the explanatory variables is generally much smaller than n.

Under the above assumptions, the maximum-likelihood (ML) estimator of the regression coefficients is the ordinary least-squares estimator, whose expression does not depend on $\boldsymbol{\Sigma}$. For all $\boldsymbol{\Sigma}$, the ML estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is indeed defined as follows: $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}$, where $\boldsymbol{S}_{xy}$ stands for the $p \times m$ sample covariance matrix between $\boldsymbol{Y}$ and $\boldsymbol{x}$.

The distribution of $\hat{\boldsymbol{\beta}}$ yet depends on the conditional dependence across response variables since the common correlation matrix of each row of $\hat{\boldsymbol{\beta}}$ is the conditional correlation matrix $\boldsymbol{R} = \boldsymbol{D}_\sigma^{-1}\boldsymbol{\Sigma}\boldsymbol{D}_\sigma^{-1}$, where D_σ is the $m \times m$ diagonal matrix whose diagonal entries are the conditional standard deviations $\sigma_j,\ j = 1, \ldots, m$. Therefore, strong conditional correlation across response variables, as for example when those response variables are densely discretized curves, is pointed out by many authors as a major cause of instability of feature selection procedures whose aim is to find the non-zero coefficients in $\boldsymbol{\beta}$.

Searching for time points where experimental variables have an effect on the hemodynamic curves can be viewed as such a feature selection issue. This can either be addressed by multiple testing procedures or by a more global estimation approach in which sparsity of $\boldsymbol{\beta}$ is accounted for by an ℓ_1–regularization of the ML procedure. In such procedures, the penalized deviance to be minimized is, up to an additive constant:

$$\mathcal{D}(\boldsymbol{\beta}, \boldsymbol{\Omega}; \kappa) = -2n \text{ trace}\big(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{S}_{yx}\big) + n \text{ trace}\big(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}\boldsymbol{S}_{xx}\big) + \kappa||\boldsymbol{\beta}||_1, \tag{2}$$

where $||.||_1$ is the sum of absolute values of terms of a matrix or a vector, $\kappa \geq 0$ is the regularization parameter and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

A Cyclic Coordinate Descent (CCD) algorithm is implemented to obtain the estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{V}; \kappa)$ minimizing $\mathcal{D}(\boldsymbol{\beta}, \boldsymbol{V}; \kappa)$ for any positive definite $m \times m$ matrix $\boldsymbol{V}$. For $\kappa = 0$, $\hat{\boldsymbol{\beta}}(\boldsymbol{V}; \kappa)$ is the same, whatever the choice of $\boldsymbol{V}$. However, this assertion is no more true for $\kappa \neq 0$.

In order to illustrate how the choice of $\boldsymbol{V}$ influences the feature selection procedure based on $\hat{\boldsymbol{\beta}}(\boldsymbol{V}; \kappa)$, this ℓ_1–regularized estimation procedure is implemented hereafter in a study design whose aim is to test for the effect of the phonological density of a word on visual recognition, using fNIRS at different locations on brain.

Phonological density study

Phonological density refers to the number of words that can be generated by replacing a phoneme in a target word with another phoneme in the same position. Chen *et al.*

(2011) investigated neurobehavioral correlates of phonological neighborhood density in skilled readers of English using fNIRS in a lexical decision task. Participants were instructed to make a speeded lexical decision response and press the response button only if they thought the stimulus formed an actual English word. The recordings of the fNIRS curves were taken over a time period from -2 seconds before stimulus on-set to 15 seconds afterwards with 200 samples per second.

A comparative study is conducted hereafter, based on data-driven simulated datasets with same dimensions and same conditional dependence matrix $\tilde{\mathbf{\Sigma}}$ as observed in the PND study introduced above. Each of the n rows of simulated dataset are independently generated from a multivariate normal distribution as in Model (1). Two scenarios are considered for the difference curve $t \mapsto \alpha_2(t)$ between the two PND conditions, all other effect curves being set to zero (see Figure 1).

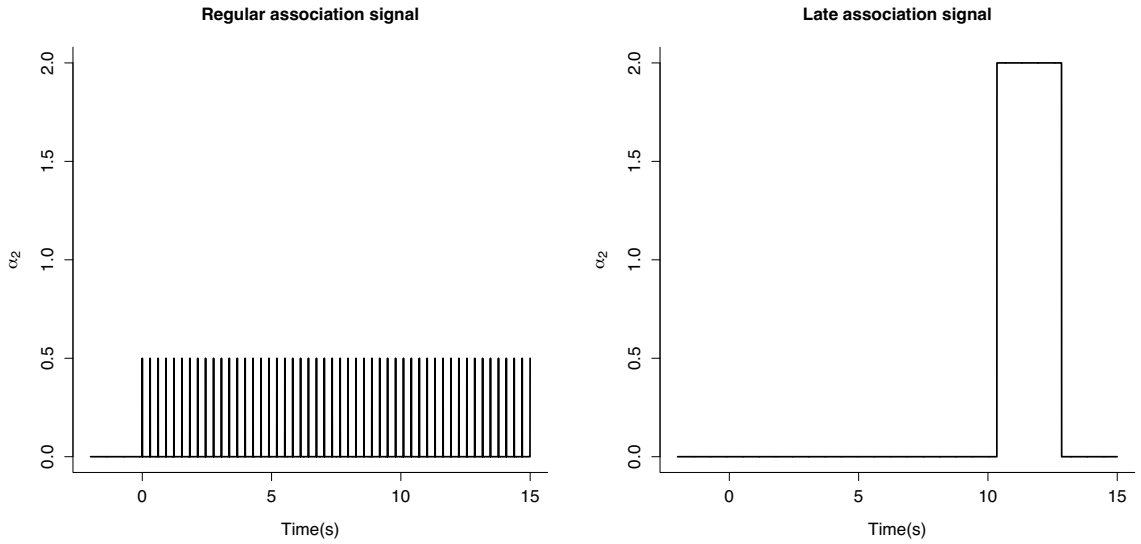


Figure 1: Difference curve $t \mapsto \alpha_2(t)$ between the two PND conditions in the two scenarios of the simulation study.

For each of 100 simulated datasets in each scenario, the penalized estimator $\hat{\boldsymbol{\alpha}}_2(\tilde{\mathbf{\Omega}}; \kappa)$ of $(\alpha_2(t_1), \ldots, \alpha_2(t_L))'$ minimizing (2), with $\tilde{\mathbf{\Omega}} = \tilde{\mathbf{\Sigma}}^{-1}$, is first calculated for a sequence of penalty parameters κ such that $-8 \leq \log \kappa \leq 2$. An alternative estimator $\hat{\boldsymbol{\alpha}}_2(\tilde{\Omega}_0; \kappa)$ is calculated, also minimizing (2) for the same sequence of penalty parameters, after replacing the complete variance-covariance matrix by the diagonal matrix $\tilde{\Omega}_0) = \boldsymbol{D}_{\tilde{\sigma}^2}$ obtained by setting all the off-diagonal terms to zero in $\tilde{\mathbf{\Sigma}}$. In other words, consistently with the model used to generate the data, the first estimator fully accounts for the conditional time dependence of the response variables whereas the second estimator ignores dependence.

Let us focus on the feature selection performance of $\hat{\boldsymbol{\alpha}}_2(\tilde{\mathbf{\Omega}}; \kappa)$ and $\hat{\boldsymbol{\alpha}}_2(\tilde{\mathbf{\Omega}}_0; \kappa)$ in the

two simulation scenarios introduced above. For each simulated dataset, the selection performance is measured by the Area under the ROC curve (AUC). Boxplots of the AUCs over 100 simulations in the two scenarios are displayed in Figure 2. It is deduced that $\hat{\boldsymbol{\alpha}}_2(\tilde{\boldsymbol{\Omega}};\kappa)$ clearly outperforms $\hat{\boldsymbol{\alpha}}_2(\tilde{\boldsymbol{\Omega}}_0;\kappa)$ in the regular association signal scenario whereas the opposite ranking of the two ℓ_1-penalized estimators is observed in the scenario of a late interval of association.

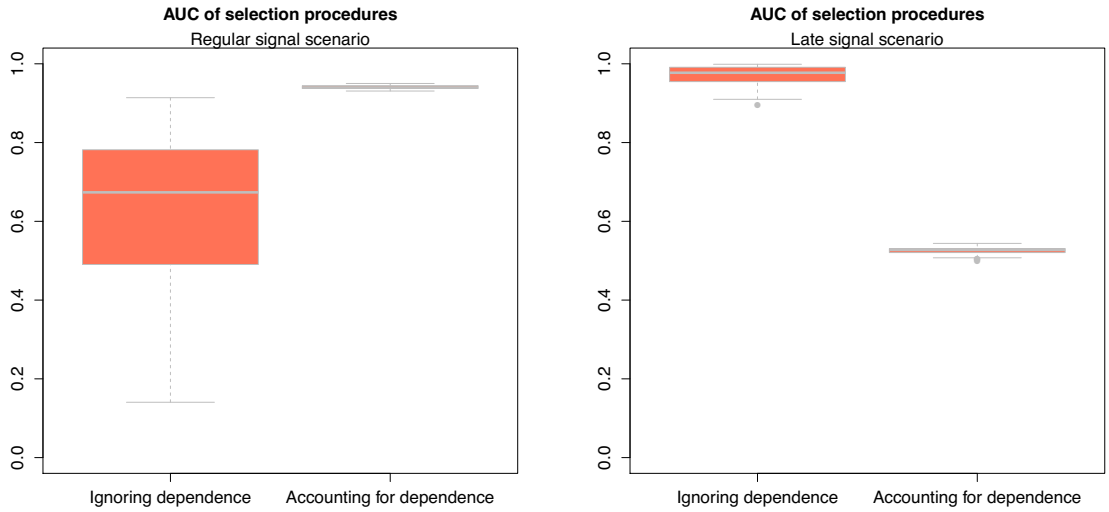


Figure 2: Boxplots of Area Under the ROC curve (AUC) over 100 simulations in each simulation scenario (Left: regular association signal, Right: late association signal) for the two ℓ_1-penalized estimators $\hat{\boldsymbol{\alpha}}_2(\tilde{\boldsymbol{\Omega}};\kappa)$ (accounting for dependence) and $\hat{\boldsymbol{\alpha}}_2(\tilde{\boldsymbol{\Omega}}_0;\kappa)$ (ignoring dependence).

The main general messages that can be deduced from the former illustrative example can be briefly stated as follows:

- For $\kappa \neq 0$, $\hat{\boldsymbol{\beta}}(\boldsymbol{V};\kappa)$ depends markedly on the choice of $\boldsymbol{V}$. Incidentally, the choice of $\boldsymbol{V}$ has a crucial impact on the feature selection performance of the ℓ_1-penalized estimation procedure.
- For some patterns of association signal, the basic choice of $\boldsymbol{V} = \boldsymbol{\Omega}_0$ consisting in ignoring dependence across response variables may lead to a much better identification of the nonzero regression parameters than $\boldsymbol{V} = \boldsymbol{\Omega}$ fully accounting for dependence.

In the following, a rank-reduced model for $\boldsymbol{\Omega}$ is introduced, offering a flexible framework for the handling of dependence through the choice of $\boldsymbol{V}$.

Rank-reduced model for conditional dependence

Hereafter, conditional dependence across response variables is assumed to have a latent q–factor structure, with $q \ll m$. In the former regression factor model, the conditional variance-covariance $\boldsymbol{\Sigma}$ of the responses given the explanatory variables in model (1) has the following decomposition: $\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{B}'$, where $\boldsymbol{\Psi}$ is a $m \times m$ diagonal matrix whose diagonal entries ψ_j^2 are positive and $\boldsymbol{B}$ is a $m \times q$ matrix.

The number q of latent factors is a tuning parameter for the handling of conditional dependence, providing intermediate strategies between ignorance of dependence, with $q = 0$, and fully accounting for it, with q as large as possible. Moreover, especially when the response variables are discretized observations of curves on a high resolution time grid, conditional dependence across responses given the explanatory variables is strong, which leads to a dense $\boldsymbol{\Sigma}$ and subsequently to a sparse inverse matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

The following new parameterization of the factor model is introduced:

$$\begin{aligned} \boldsymbol{\varphi} &= \boldsymbol{\Psi}^{-\frac{1}{2}}, \\ \boldsymbol{\theta} &= \tilde{\boldsymbol{B}}(\boldsymbol{I}_q + \tilde{\boldsymbol{B}}'\tilde{\boldsymbol{B}})^{-\frac{1}{2}}, \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{B}} = \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{B}$ is the $m \times q$ matrix of normalized loadings. Using Woodbury's identity, it is straightforwardly checked that $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \boldsymbol{\varphi}(\boldsymbol{I}_m - \boldsymbol{\theta}\boldsymbol{\theta}')\boldsymbol{\varphi}$ also has a q-factor structure.

In order to account simultaneously on the sparsity of the association signal and $\boldsymbol{\Omega}$, we propose to estimate $\boldsymbol{\beta}$ by minimization of the following doubly penalized deviance:

$$\mathcal{D}(\boldsymbol{\beta}, \kappa_1, \kappa_2) = -2n \text{ trace}\big(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{S}_{yx}\big) + n \text{ trace}\big(\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}\boldsymbol{S}_{xx}\big) + \kappa_1||\boldsymbol{\beta}||_1 + \kappa_2||\boldsymbol{\theta}||_r^r, \tag{4}$$

where $r = 1$ or 2, $||.||_2^2$ is the sum of squared values of terms of a matrix or a vector, $\kappa_1 \geq 0$ and $\kappa_2 \geq 0$ are regularization parameters.

The presentation will demonstrate how the minimization of (4) can lead to optimal strategies for handing dependence in the feature selection issue introduced for functional data.

References

Causeur, D., Sheu, C. F., Perthame, E. and Rufini, F (2020). A functional generalized F-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*. 76(1), 246-256.

Chen, H.-C., Vaid, J., Boas, D. A., and Bortfeld, H. (2011). Examining the phonological neighborhood density effect using near infrared spectroscopy. *Human Brain Mapping*, 32(9), 1363–1370.

Un modèle à blocs stochastiques pour les réseaux multiniveaux

Saint-Clair Chabert-Liddell [†,1] & Pierre Barbillon [†,2] & Sophie Donnet [†,3] & Emmanuel Lazega [⋆,4]

[†] *UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, 75005, Paris, France*

[⋆] *Institut d'Études Politiques de Paris, France*

[1] *saint-clair.chabert-liddell@agroparistech.fr*

[2] *pierre.barbillon@agroparistech.fr*

[3] *sophie.donnet@inrae.fr*

[4] *emmanuel.lazega@sciencespo.fr*

Résumé. Nous définissons un réseau multiniveau comme la jonction de deux réseaux d'interaction, l'un représentant les interactions entre individus et l'autre les interactions entre organisations. Ces niveaux sont reliés par une relation d'appartenance, chaque individu appartenant à une unique organisation. Le modèle à blocs stochastiques (SBM) est un modèle à variables latentes qui permet de modéliser l'hétérogénéité des connexions d'un réseau en classifiant les nœuds suivants leurs profils de connectivité. Nous étendons le SBM au cas des réseaux multiniveaux et prouvons l'identifiabilité de ce nouveau modèle. Les paramètres de notre modèle sont estimés par des méthodes variationnelles (algorithme VEM) et nous développons un critère de vraisemblance complète intégrée (ICL) pour sélectionner non seulement le nombre de blocs mais également pour décider de la dépendance ou non entre les structures des deux niveaux. Nous justifions via des simulations l'intêret de notre approche ainsi que la robustesse de notre méthode d'estimation de paramètres et de notre critère de sélection de modèle. Nous appliquons notre modèle sur des données collectées lors d'un salon audiovisuel. Le niveau inter-organisationnel représente le réseau des relations économiques entre entreprises et le niveau inter-individuel celui des relations informelles entre leurs représentants sur ce salon.

Mots-clés. Modèle à variables latentes, modèle hiérarchique, réseaux sociaux, inférence variationnelle

Abstract. We define a multilevel network as the junction of two interaction networks, one level representing the interactions between individuals and the other one the interactions between organizations. The levels are linked by an affiliation relationship, each individual belonging to a unique organization. We design a Stochastic block model (SBM) suited to multilevel networks. SBM is a latent variable model for networks, where the connections between nodes depend on a latent clustering (blocks), thus modeling some connection heterogeneity. We prove the identifiability of our model. The parameters of the model are estimated with a variational EM algorithm. An Integrated Completed

Likelihood criterion is developed not only to select the number of blocks but also to detect whether the two levels (individuals and organizations) are dependent or not. In a comprehensive simulation study, we exhibit the benefit of considering our approach, illustrate the robustness of our parameter estimation and highlight the reliability of our model selection criterion. Our approach is applied on a sociological dataset collected during a television program trade fair. The inter-organizational level is the economic network between companies and the inter-individual level is the informal network between their representatives.

Keywords. Latent variable model, Hierarchical modeling, Social networks, Variational inference

1 A multilevel stochastic block model (MLVSBM)

In what follows, a multilevel network is defined as the collection of an inter-individual network, an inter-organizational network and the affiliation of the individuals to the organizations. Besides, we assume that the individuals belong to a unique organization. All the results are given for undirected networks.

Let us consider n_I individuals involved in n_O organizations. We encode the networks into adjacency matrices as follows. Let X^I be the binary $n_I \times n_I$ matrix representing the inter-individual network. X^I is such that : $\forall (i, i') \in \{1, \ldots, n_I\}^2$:

$$X^I_{ii'} = \begin{cases} 1 & \text{if there is an interaction from individual } i \text{ to individual } i', \\ 0 & \text{otherwise.} \end{cases}$$

X^O is the binary $n_O \times n_O$ matrix representing the inter-organizational network, $\forall (j, j') \in \{1, \ldots, n_O\}^2$:

$$X^O_{jj'} = \begin{cases} 1 & \text{if there is an interaction from organization } j \text{ to organization } j', \\ 0 & \text{otherwise.} \end{cases}$$

Let A be the affiliation matrix. A is a $n_I \times n_O$ matrix such that:

$$A_{ij} = \begin{cases} 1 & \text{if individual i belongs to organization j,} \\ 0 & \text{otherwise} \end{cases}.$$

A is such that $\forall i = 1, \ldots, n_I$, $\sum_{j=1}^{n_O} A_{ij} = 1$ since we assume that any individual belongs to a unique organization.

We propose a joint modeling of the inter-individual and inter-organizational networks based on an extension of the stochastic block model (SBM; Snijders and Nowicki, 1997).

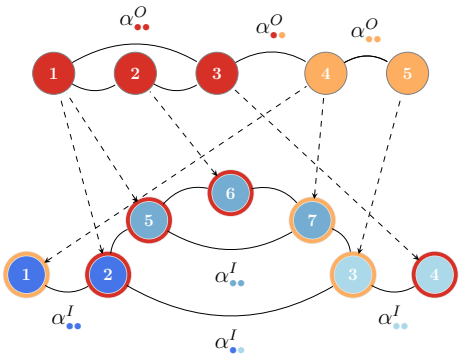


Figure 1: Representation of a multilevel network with inter-organizational level on the top and inter-individual level on the bottom. .

More precisely, assume that the n_O organizations are divided into Q_O blocks and that the n_I individuals are divided into Q_I blocks. Let $Z^O = (Z_1^O, \ldots, Z_{n_O}^O)$ and $Z^I = (Z_1^I, \ldots, Z_{n_I}^I)$ be such that $Z_j^O = l$ if organization j belongs to cluster l ($l \in \{1, \ldots, Q_O\}$) and $Z_i^I = k$ if individual i belongs to cluster k ($k \in \{1, \ldots, Q_I\}$).

Given these clusterings, we assume that the interactions between organizations and interactions between the individuals are independent and distributed as follows:

$$\begin{aligned} \mathbb{P}(X_{jj'}^O = 1 | Z_j^O, Z_{j'}^O) &= \alpha^O_{Z_j^O Z_{j'}^O} \\ \mathbb{P}(X_{ii'}^I = 1 | Z_i^I, Z_{i'}^I) &= \alpha^I_{Z_i^I Z_{i'}^I}. \end{aligned} \tag{1}$$

As a consequence, the blocks gather nodes sharing the same profiles of connectivity.
In order to take into account the fact that organizations may shape the individual behaviors, we assume that the memberships of the individuals (Z^I) depend on the cluster of the organizations (Z^O) they are affiliated to. More precisely, we set:

$$\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) = \gamma_{kZ_j^O} \quad \forall i \in \{1, \ldots, n_I\} \quad \forall k \in \{1, \ldots, Q_I\} \tag{2}$$

where γ is a $Q_I \times Q_O$ matrix such that $\sum_{k=1}^{Q_I} \gamma_{kl} = 1$ for any $l \in \{1, \ldots, Q_O\}$. The (Z_j^O) are assumed to be independent random variables distributed as

$$\mathbb{P}(Z_j^O = l) = \pi_l^O, \qquad \forall j \in \{1, \ldots, n_O\} \quad \forall l \in \{1, \ldots, Q_O\} \tag{3}$$

with $\sum_{l=1}^{Q_O} \pi_l^O = 1$.
A small multilevel network is depicted in Figure 1.

Independence. We derive conditions for the structural independence between levels in term of parameters equality.

Proposition 1. *In the MLVSBM, the two following properties are equivalent: [1.]: Z^I is independent on Z^O, [2.]: $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \ldots, Q_O\}$ and imply that: [3.]: X^I and X^O are independent.*

Identifiability. We adapt the proof given in Celisse et al. (2012) to obtain the identifiability of the MLVSBM.

Proposition 2. *The MLVSBM is identifiable up to label switching under the following assumptions:*

A1. All coefficients of $\alpha^I \cdot \gamma \cdot \pi^O$ *are distinct and all coefficients of* $\alpha^O \cdot \pi^O$ *are distinct.*

A2. $n_I \geq 2Q_I$ *and* $n_O \geq \max(2Q_O, Q_O + Q_I - 1)$.

A3. At least $2Q_I$ *organizations contain one individual or more.*

Likelihood. From Equations (1), (2) and (3), we derive the complete log-likelihood for a directed MLVSBM where θ denotes all the model parameters:

$$\begin{aligned}
&\log \ell_\theta \left(X^I, X^O, Z^I, Z^O | A\right) \\
&\quad = \log \ell_{\pi^O}(Z^O) + \log \ell_\gamma(Z^I | Z^O, A) + \log \ell_{\alpha^I}(X^I | Z^I) + \log \ell_{\alpha^O}(X^O | Z^O) \\
&\quad = \sum_{j,l} \mathbb{1}_{Z_j^O = l} \log \pi_l^O + \sum_{i,k} \mathbb{1}_{Z_i^I = k} \sum_{j,l} A_{ij} \mathbb{1}_{Z_j^O = l} \log \gamma_{kl} \qquad (4) \\
&\quad + \frac{1}{2} \sum_{i' \neq i} \sum_{k,k'} \mathbb{1}_{Z_i^I = k} \mathbb{1}_{Z_{i'}^I = k'} \log \phi(X_{ii'}^I, \alpha_{kk'}^I) + \frac{1}{2} \sum_{j' \neq j} \sum_{l,l'} \mathbb{1}_{Z_j^O = l} \mathbb{1}_{Z_{j'}^O = l'} \log \phi(X_{jj'}^O, \alpha_{ll'}^O),
\end{aligned}$$

where $\phi(x, a) = a^x (1 - a)^{1-x}$.

2 Statistical Inference, Simulations and Applications

Variational method for maximum likelihood estimation Due to the latent variables, the estimation of the parameters is a complex task. The likelihood of $\mathbf{X} = \{X^I, X^O\}$ $\ell_\theta(\mathbf{X}|A)$ is obtained by integrating out the latent variables $\mathbf{Z} = \{Z^I, Z^O\}$ in the complete data likelihood (4). However, this calculus becomes not computationally tractable as the number of nodes and blocks grow.

The Expectation-Maximisation algorithm (EM) (Dempster et al., 1977) is a popular solution to maximize the likelihood of models with latent variables, but it requires the computation of $\mathbb{P}(\mathbf{Z}|\mathbf{X}, A)$ which is also not tractable in our case. Hence, as Daudin et al. (2008) did for the SBM, we resort to a variational version of the EM algorithm.

The variational EM algorithm aims to maximize a lower bound of $\log \ell_\theta(\mathbf{X}|A)$ by iterating two steps. Step VE maximizes the lower bound with respect to the parameters of an approximate distribution of $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)$. Step M maximizes the lower bound with respect to the model parameters θ.

Model selection Following Biernacki et al. (2000) and Daudin et al. (2008), we propose an ad-hoc version of the Integrated Complete Likelihood (ICL) criterion to choose the number of blocks. It is an asymptotic approximation of the complete likelihood integrated over its parameters and latent variables given for the MLVSBM by:

$$ICL(Q_I, Q_O) = \log \ell_{\hat{\theta}}(X^I, X^O, \widehat{Z^I}, \widehat{Z^O} | A, Q_I, Q_O) - pen(Q_I, Q_O),$$

where

$$\begin{aligned} pen(Q_I, Q_O) = & \frac{1}{2} \frac{Q_I(Q_I+1)}{2} \log \frac{n_I(n_I-1)}{2} + \frac{Q_O(Q_I-1)}{2} \log n_I + \\ & \frac{1}{2} \frac{Q_O(Q_O+1)}{2} \log \frac{n_O(n_O-1)}{2} + \frac{Q_O-1}{2} \log n_O \end{aligned}$$

where $\widehat{Z^O}$ and $\widehat{Z^I}$ are the imputed latent variables using the maximum a posteriori (MAP) of $\mathbb{P}_{\hat{\theta}}(\mathbf{Z}|\mathbf{X}, A; Q_I, Q_O)$.

We also use the ICL criterion to assess whether the two levels of interactions are independent or not by comparing the ICL of the MLVSBM with the sum of the ICL of two independent SBMs, one for each level.

We provide a stepwise procedure for model selection which seeks for the optimal number of blocks at a reasonable cost. As a by-product, it states whether the two levels are independent or not. To simulate and infer the MLVSBM, we developed our own `R` package available at https://chabert-liddell.github.io/MLVSBM/.

Simulation studies We study the performances of the inference procedure for the MLVSBM including the ability to recover blocks, the selection of the numbers of blocks and the independence detection.

For $Q_I = Q_O = 3$, on three standard topologies and for different values of density, we set $\gamma_{kk} = \delta$ and $\gamma_{kk'} = .5(1-\delta)$ for $k \neq k'$, $k, k' \in \{1, 2, 3\}$. δ is a parameter for the strength of the dependence between levels ranging from 0 to 1. When $\delta = 1/3$ the levels are independent.

We compare the SBM and the MLVSBM in their abilities to recover the clusters of the inter-individual level (Figure 2 A), the true number of blocks ($Q_I = 3$) and to capture the interdependence between the two levels (Figure 2 B and C).

Application We apply our model to a sociological dataset collected during a television program trade fair. The inter-organizational level is the economic network between companies and the inter-individual level is the informal network between their representatives. Our results exhibit the structural interdependence between the two levels and in particular the heterogeneity of individuals who replicate to different extent their organizations's ties.

A preprint is available at `https://hal.archives-ouvertes.fr/hal-02353711v1`.

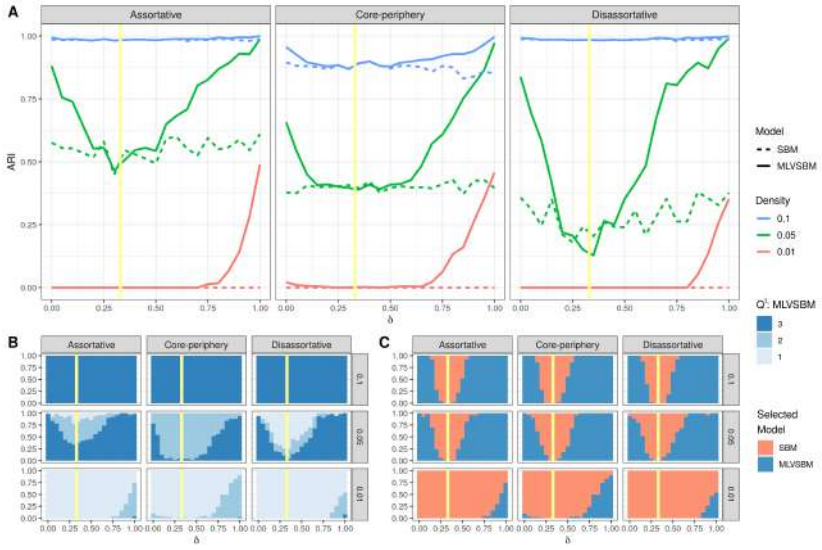


Figure 2: Clustering and model selection accuracy for 3 different topologies and densities on the inter-individual level, as function of δ. The yellow vertical lines corresponds to $\delta = 1/3$ ensuring the independence between the two levels. **A**: ARI (Adjusted Rand Index) for the inter-individual level, comparing the MLVSBM with two independent SBMs. **B**: Number of blocks for the inter-individual level chosen by the ICL criterion for the MLVSBM. **C**: Model selected by the ICL for the inter-level dependence (either MLVSBM or two independent SBMs).

References

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*(7), 719–725.

Celisse, A., J.-J. Daudin, and L. Pierre (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics 6*, 1847–1899.

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and computing 18*(2), 173–183.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Snijders, T. A. and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification 14*(1), 75–100.

Reconstruction of motion signals with curvature and torsion

Perrine Chassat [1] & Nicolas Brunel [2,3] & Juhyun Park[2]

[1] *Université Paris Saclay, CNRS, Univ. Evry, Laboratoire de Mathématiques et Modélisation d'Evry, perrine.chassat@gmail.com*
[2] *Université Paris Saclay, CNRS, ENSIIE, Laboratoire de Mathématiques et Modélisation d'Evry, nicolas.brunel@ensiie.fr, juhyun.park@ensiie.fr*
[3] *Quantmetry*

Résumé. Cette communication propose un algorithme d'estimation non paramétrique de la courbure, de la torsion et du trièdre de Frenet d'une courbe dans $\mathbb{R}^3$, dans le but de proposer une analyse statistique de données du mouvement acquis par Motion Capture, pour la détection et la caractérisation de primitives du mouvement. Nous utilisons pour cela une représentation géométrique des courbes par le biais des équations différentielles ordinaires de Frenet-Serret et nous formulons le problème d'estimation par une régression pénalisée et développons un algorithme efficace, dont nous évaluons la qualité par la reconstruction de la trajectoire. Ces travaux sont motivés par l'analyse du mouvement d'une main dans le contexte de la langue des signes, acquis par Motion Capture par l'entreprise MOCAPLAB[1].

Mots-clés. Langue des signes, Capture de mouvement, Analyse de trajectoires cinématiques, Analyse de la forme, Analyse de données fonctionnelles, Estimation de la courbure, Repère de Frenet-Serret

Abstract. This communication proposes a nonparametric estimation algorithm for the curvature, torsion and Frenet path of a curve in $\mathbb{R}^3$, with the aim to propose a statistical analysis of motion data acquired by Motion Capture, for the detection and characterization of movement primitives. For this purpose, we directly use a geometric representation of the curves through the Frenet-Serret ordinary differential equations. We formulate the estimation problem in a penalized regression and develop an efficient algorithm, whose quality is evaluated by reconstruction of the trajectory. We apply our method to real data of a hand motion in the context of sign language, acquired by Motion Capture from the company MOCAPLAB.

Keywords. Sign language, Motion capture, Kinematic trajectory analysis, Shape analysis, Functional data analysis, Curvature estimation, Frenet-Serret frame

1 Introduction

Among the many fields of exploration of motion capture is the very specific field of sign language involving movements of the body, hands, fingers, face and eyes and achieve a

[1] https://www.mocaplab.com/fr/

capacity for expression as rich and structured as that offered by speech [5]. These types of movement are interesting for us as they are particularly meaningful but are difficult to characterize without specific knowledge in the field. Our idea in treating the "motion" signals, in collaboration with the company MOCAPLAB, specialized in Motion Capture, is to incorporate current mathematical and statistical tools to extract "primitives" specific to the nature of the signals studied.

From the perspective of motor control theories, the axis of the analysis envisaged should be able to exploit the link between speed and motion geometry [4]. The starting point of our methodology consists in the estimation of trajectories resulting from motion capture and the characterization of the geometry and kinematics by appropriate functional representations. This simultaneous analysis is particularly possible in the trajectory of a point particle, using the Frenet-Serret frame [1, 2].

Our approach is to focus on estimating curvature and torsion [6, 7]. Indeed, the geometry of the trajectories of movement have physical significance: curvature and torsion characterize this geometry and can provide insightful summaries of kinetic curves. Discovering and exploiting these relationships require a good estimation of the functional parameters such as curvature and Frenet paths. This is a challenging statistical task as curvature and torsion depend on higher order derivatives and their estimation from real data (even with a low noise) can be very unstable.

Our work is motivated by the new development in [8], which offers a unified framework for the estimation of these functional quantities in the setting of functional data. A motivation of this type of analysis is to exploit the link between the curvilinear speed and the geometry of the trajectories (typically curvatures).Although the previous work is designed for multiple trajectories as an extension of functional data, it is still applicable to single curves and here we adopt the geometric framework for the analysis of single trajectories. Arguably, this is more challenging as the benefit of borrowing information from multiple curves is missing in the single curve setting.

Under this framework, the problem of estimating curvature and torsion can be treated as estimation of an ordinary differential equation in a Lie group. We develop a new algorithm for estimation by considering the perturbed Frenet-Serret ODE and solving the optimal control problem with a computationally fast Kalman filter [3]. The quality of the estimates is evaluated based on the quality of reconstruction of the trajectory by solving the Frenet-Serret ordinary differential equation. In particular, we show that our proposed method dominates the straightforward estimates of curvature and torsion based on the extrinsic formulas.

2 Frenet-Serret representation of curves

We consider curves in $\mathbb{R}^3$ defined as functions $x : [0, T] \mapsto \mathbb{R}^3$. In order to avoid some technical difficulties, we assume that the curves are regular, i.e they are of class $\mathcal{C}^r$, $r \geq 3$ (w.r.t time t), the time derivative $\dot{x}(t)$ never vanishes on $[0, T]$ and the curves never intersect themselves. For a curve x, the shape of the curve $X : [0, L] \mapsto \mathbb{R}^3$ is the image

of the function x, which satisfies $x(t) = X(s(t))$, where $s(t) = \int_0^t ||\dot{x}(u)||du$ the arclength, $s(T) = L$ is the total length of the curve $X = \{x(t), t \in [0,T]\}$ and $\dot{s}(t) = ||\dot{x}(t)||$ the curvilinear speed. The tangent vector is defined as $T(s) = X'(s)$ and the curvature as $s \mapsto \kappa(s) = ||T'(s)||$. The moving frame is defined with the addition of Normal vector $N(s) = \frac{1}{\kappa(s)}T'(s)$ and the bi-normal vector $B(s) = T(s) \times N(s)$ to T. The matrix $Q(s) = [T(s)|N(s)|B(s)]$ is then an orthonormal frame which can be obtained by Gram-Schmidt orthonormalisation of the frame $[X'(s)|X''(s)|X'''(s)]$. This Frenet frame is a solution of the following ODE

$$\begin{cases} T'(s) = \kappa(s)N(s) \\ N'(s) = -\kappa(s)T(s) + \tau(s)B(s) \\ B'(s) = -\tau(s)N(s) \end{cases}$$

where the functions $s \mapsto \kappa(s), \tau(s)$ are respectively the curvature and torsion. They are geometric invariant of the curve, independent of the parametrization of a curve X and invariant under the action of rigid (Euclidean) motions. They can be directly defined with extrinsic formulas as

$$\kappa(s(t)) = \frac{||\dot{x}(t) \times \ddot{x}(t)||_2}{||\dot{x}||_2^3}, \qquad \tau(s(t)) = \frac{\langle \dot{x}(t) \times \ddot{x}(t), \dddot{x}(t) \rangle}{||\dot{x}(t) \times \ddot{x}(t)||_2^2} \tag{1}$$

An alternative interpretation of this Frenet-Serret formula is that it defines an ODE in the Lie group $SO(3)$ as

$$\dot{Q}(s) = Q(s)A_\theta(s), \qquad A_\theta(s) = \begin{bmatrix} 0 & -\kappa(s) & 0 \\ \kappa(s) & 0 & -\tau(s) \\ 0 & \tau(s) & 0 \end{bmatrix} \tag{2}$$

with the generalised curvature $\theta = (\kappa, \tau)$ and the initial condition $Q(0) = Q_0$.

3 Frenet paths and curvatures estimation algorithm

We assume that we have discrete and noisy observations $Y_j, j = 1, \ldots, n$ of the Frenet path Q at s_j where $0 = s_1 < s_2 ... < s_n = 1$. To take into account the noise, we consider the following statistical model $Y_j = Q(s_j)\exp(W_j)$ where W_j are random matrices in $SO(p)$. Our problem is to recover the Frenet path $s \mapsto Q(s)$ from the noisy observations Y_j and to derive the best parameters θ that fit the data under the dynamical model $\dot{Q}(s) = Q(s)A_\theta(s)$. So for all j, and for all t in $[0,1]$, we should have approximately $\phi_\theta(t - s_j, s_j, Y_j) \approx Q(t)$, where the flow $t \mapsto \phi_\theta(t, s, M)$ is the solution of the Frenet-Serret ODE satisfying $Q(s) = M$.

In order to fit the model, [8] proposes a simultaneous estimation of Q and θ by the minimization of the following criterion:

$$\mathcal{J}_{h,\lambda}(\theta, \mathbf{Q}; \mathbf{Y}) = \frac{1}{n}\sum_{j=1}^{n} \int_0^1 K_h(t - s_j)|| \log\left(Q(t)^T \phi_\theta(t - s_j, s_j, Y_j)\right) ||_F^2 dt + \lambda \int_0^1 ||\theta''(t)||^2 dt$$

where $K_h(\cdot) = (1/h)K(\cdot/h)$ is a weight function $K(t) \geq 0, \int K = 1$ introduced for taking into account the increasing uncertainty for distant points.

In order to get a computable criterion, the integral is discretized on the grid $0 = t_1 < t_2 < ... < t_T = 1$ and we replace the exact flow with a first order approximation to the Magnus expansion.

$$\begin{aligned}\tilde{\mathcal{J}}_{h,\lambda}(\theta, \mathbf{Q}; \mathbf{Y}) &= \frac{1}{nT} \sum_{j,q=1}^{n,T} K_h(u_{jq}) u_{jq}^2 ||A_\theta(v_{jq}) - \tilde{R}_{jq}||_F^2 dt + \lambda \int_0^1 ||\theta''(t)||^2 dt \\ &= \sum_{j,q=1}^{n,T} \omega_{jq}(\kappa(v_{jq}) - r_{jq}^1)^2 + \sum_{j,q=1}^{n,T} \omega_{jq}(\tau(v_{jq}) - r_{jq}^2)^2 + \lambda \int_0^1 ||\theta''(t)||^2 dt\end{aligned}$$

with $u_{jq} = t_q - s_j$, $v_{jq} = \frac{t_q + s_j}{2}$, $\omega_{jq} = \frac{1}{nT} K_h(u_{jq}) u_{jq}^2$ and $\tilde{R}_{jq} = -\frac{1}{u_{jq}} \log(Q(t_q)^T Y_j)$.

Then, the estimation of θ from a given path Q is done by solving the optimization problem

$$\tilde{\theta}_{h,\lambda}(., \mathbf{Q}) = \min_\theta \tilde{\mathcal{J}}_{h,\lambda}(\theta, \mathbf{Q}; \mathbf{Y})$$

which is made by two independent weighted smoothing splines. This estimation must be alternated with that of the Frenet path Q to be able to compute the pseudo-observation r_{jq}^1, r_{jq}^2. The corresponding final estimates for θ and Q are denoted respectively $\hat{\theta}$ and $\hat{\boldsymbol{Q}}$. We propose in this work an alternative way of estimating the Frenet path Q compared to the algorithm in [8]. Instead of computing the Karcher mean for estimating Q, we adapt the framework developed in [3] to the Frenet-Serret ODE. We consider the perturbed Frenet-Serret ODE, $\dot{Q}(s) = Q(s)A_\theta(s) + U(s)B(s)$ with $U \in SO(p)$ and the fitting problem can be cast into a tracking problem, where the smallest optimal U must be found. This control problem can be solved very fast if we consider an appropriate Linear-Quadratic (LQ) problem.

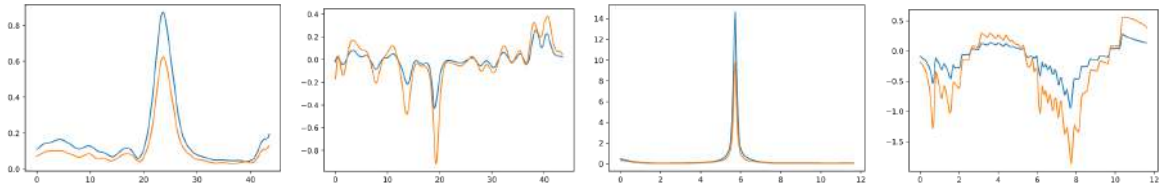
4 Application to real data

We consider the trajectory of a point in the middle of the palm of a man's hand when he speaks sign language, acquired by Motion Capture. The trajectory of this point on the right hand when the entire sentences, *"Temoignage : Evan, un jeune autiste qui prepare son avenir"*, is signed, is visible in Figure 1.

We consider two pieces of this trajectory. We pre-process the data to create an arclength parametrized data X_i defined on $[0, L_i], i = 1, 2$. We consider a grid with constant bandwidth. The noisy raw Frenet paths (Y_i) are obtained from Gram-Schmidt orthonormalization from derivatives estimated by constrained local polynomial smoothing with order 5.

Figure 2 shows the estimates of curvature and torsion by the algorithm described in section 3, as well as the one estimated from extrinsic formulas for comparison. The quality of our estimates is illustrated in Figure 3 which displays the reconstructions of the trajectory from them (ODE resolution and then integration of the tangent vector).

Figure 1: Example of set up for sign language movement capture in MOCAPLAB on the right and entire trajectory of the right hand when a sentence is signed on the left.



(a) curvature X_1 (b) torsion X_1 (c) curvature X_2 (d) torsion X_2

Figure 2: In blue estimates from extrinsic formulas, in orange estimates from our algorithm.

(a) X_1 (b) X_2

Figure 3: Reconstructions of the shape. In green: raw X. In blue: reconstruction from θ estimated using extrinsic formulas. In orange: reconstruction from $\hat{\theta}$, estimated with our algorithm. In purple: Reconstruction from our estimated Frenet path $\hat{\boldsymbol{Q}}$.

.

The hyperparameters λ_κ, λ_τ, h, (used for θ and smoothing) and λ_Q (used for tracking Q) are obtained with Bayesian optimization. The computation time depends on several parameters, such as the number of points, the number of nodes chosen to smooth curvature and torsion, etc. With the same set of parameters, the algorithm converges about 3 times faster with this alternative method for estimating the Frenet path than with the method present in [8].

This work proposes an alternative method for estimating the Frenet paths following the framework introduced in [8]. The algorithm proposed has the main advantage of being faster and more efficient on the studied data. Moreover, its application on a single trajectory acquired by motion capture shows that it offers a robust and reliable estimation of the curvature, the torsion, better than the estimation by extrinsic formulas (as shown in Figure 3: extrinsic formulas does not pass the Frenet-Serret test). These robust estimates give access to a reliable estimation of the geometry of the curve, which plays a prominent role in motion data. Such tools might enable a deeper statistical analysis of motion data and interplay between geometry, speed and sign language.

References

[1] Brunel, N. and Park, J. *Removing phase variability to extract a mean shape for juggling trajectories.* Electronic Journal of Statistics 8.2, 2014.

[2] Brunel, N. and Park, J. *The Frenet-Serret framework for aligning geometric curves.* International Conference on Geometric Science of Information. Springer, Cham, 2019.

[3] Clairon, Q. and Brunel, N. *Tracking for parameter and state estimation in possibly misspecified partially observed linear Ordinary Differential Equations* Journal of Statistical Planning and Inference. 2019.

[4] Flash, Tamar, and al. *Motor compositionality and timing: combined geometrical and optimization approaches.* Biomechanics of Anthropomorphic Systems. Springer, Cham, 2019.

[5] Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun R. and Turki, A. *Interactive editing in French Sign Language dedicated to virtual signers: requirements and challenges* Universal Access in the Information Society, Springer Verlag, 2016.

[6] Kim, K.-R., P. Kim, J.-Y. Koo, and M. Pierrynowski *Frenet-serret and the estimation of curvature and torsion.* IEEE Journal of Selected Topics in Signal Processing 7(4), 2013.

[7] Lewiner, T., J. Gomes, H. Lopes, and M. Craizer. *Curvature and torsion estimators based on parametric curve fitting.* Computers and Graphics 29(5), 2005.

[8] Park, J. and Brunel, N. *Mean curvature and mean shape for multivariate functional data under Frenet-Serret framework.* arXiv:1910.12049, 2020.

Filtrage adaptatif de signaux définis sur des graphes de grande taille

Elie Chedemail [1,2] & Basile de Loynes[2] & Fabien Navarro[2] & Baptiste Olivier[1]

[1] *Orange Labs, France*
{*elie.chedemail, baptiste.olivier*}*@orange.com*
[2] *ENSAI, Campus de Ker Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz*
{*basile.deloynes, fabien.navarro*}*@ensai.fr*

Résumé. L'analyse de signaux sur les graphes s'attache à étendre la théorie (analyse de Fourier) et les méthodologies (filtrage notamment) du champ classique à des signaux définis sur les noeuds d'un graphe. De plus en plus populaire de par la flexibilité de la structure de graphe, ce champ de recherche trouve de nombreux domaines d'applications (réseaux de télécommunications, réseaux sociaux, chimie organique, neurologie, apprentissage profond). L'approche mise en oeuvre consiste à appliquer une procédure de seuillage dans un domaine transformé bien choisi dans lequel une représentation parcimonieuse du signal est présumée. La calibration des seuils est obtenue en minimisant l'estimateur sans biais du risque dû originellement à Stein et adapté à la transformation choisie. Le coeur de cette communication est de proposer une approche permettant le passage à l'échelle des grands graphes à l'aide d'une approximation par polynômes de Chebyshev du calcul fonctionnel. La mise en oeuvre de cette approche est illustrée numériquement sur un grand graphe (dépassant le million de noeuds).

Mots-clés. Estimateur de Stein sans biais du risque, traitement du signal sur les graphes, estimation par Monte Carlo

Abstract. Graph Signal Analysis focuses on extending the theory (Fourier analysis) and methodologies (such as filtering) of the classical field to signals defined on the vertices of a graph. Increasingly popular because of the flexibility of the underlying structure, this research area can be applied in many context such as telecommunications networks, social networks, organic chemistry, neurology or deep learning. The approach implemented in the sequel consists in applying a thresholding procedure in a well-chosen transformed domain in which the signal is presumed sparsely represented. The threshold calibration is obtained by minimizing the unbiased estimator of the risk originally due to Stein and adapted to the chosen transformation. The core of this paper is to propose an approach that scales to large graphs using a Chebyshev polynomial approximation of the functional computation. The implementation of this approach is illustrated numerically on a large graph (exceeding one million nodes).

Keywords. Stein Unbiased Risk Estimation, Graph Signal Processing, Monte Carlo estimation

1 Introduction

Les données acquises à partir de systèmes interactifs à grande échelle, tels que les réseaux informatiques, écologiques, sociaux, financiers ou biologiques, sont de plus en plus répandues et accessibles. Au sein de l'apprentissage statistique moderne, la représentation, le traitement ou l'analyse efficace de ces données structurées à grande échelle, telles que les graphes ou les réseaux, sont quelques-uns des problèmes clés (*cf.* Nickel et al. (2015); Bronstein et al. (2017)). Le domaine émergent du traitement des signaux sur graphes met en évidence les liens entre les domaines que sont le traitement du signal et de la théorie spectrale des graphes (par exemple Shuman et al. (2013); Ortega et al. (2018) pour l'illustration de telles interactions). Le rapide développement de cette thématique est illustré par la revue récente Dong et al. (2020) dans laquelle sont en outre évoquées les perspectives de cette thématique ainsi que son rôle dans certaines des premières conceptions des architectures de réseaux neuronaux sur graphes (GNN).

Basé sur une analogie entre le laplacien d'un graphe et l'opérateur de Laplace-Beltrami, une notion de transformée en ondelettes dans le contexte des graphes a été proposées dans Hammond et al. (2011). Cette construction est par ailleurs étroitement liée à celle plus générale proposée dans Coifman and Maggioni (2006) qui s'applique notamment au cas des variétés différentiables. Une construction légèrement différente proposée dans Göbel et al. (2018) permet de définir une transformée multi-échelle semi-orthogonale de sorte que l'énergie du signal est préservée dans le domaine transformé, ouvrant ainsi la porte à des méthodes de débruitages numériquement efficaces. Plus récemment, dans de Loynes et al. (2021), une méthode de sélection automatique du paramètre de seuillage a été introduite en adaptant l'estimateur sans biais du risque de Stein (SURE) à ce type de transformées semi-orthogonales. Par construction, une telle approche nécessite la diagonalisation du laplacien ce qui est rédhibitoire en grande dimension. Les difficultés à lever se situent à la fois au moment de la transformation et lors de la phase d'optimisation du SURE dont le terme de divergence fait apparaître des poids caractéristiques de la transformée. La première difficulté a déjà été traitée dans Hammond et al. (2011) en proposant une transformée rapide à base d'approximations par polynômes de Chebyshev. Le deuxième écueil, quant à lui, sera résolu par Monte Carlo en profitant de l'interprétation en terme de covariance des poids. La faisabilité d'une telle stratégie est illustrée par une comparaison numérique avec le DFS Fussed Lasso introduit dans Padilla et al. (2017). Les résultats numériques corroborent ceux de de Loynes et al. (2021) sur la pertinence de l'analyse multi-échelle face aux méthodes de pénalisations.

2 Débruitage de signaux sur un graphe

Dans toute la suite, on considère $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ un graphe non orienté composé d'un ensemble $\mathcal{V}$ de n sommets, d'un ensemble $\mathcal{E}$ d'arêtes pondérées et d'une matrice de poids W. La matrice laplacienne est donnée par $\mathcal{L} = D - W$ avec D la matrice diagonale

des degrés ($D_{i,i} = \sum_{j \in \mathcal{V}} W_{i,j}$). Par ailleurs, on considère un signal $f \in \mathbb{R}^n$ défini sur les sommets ainsi que le modèle additif de débruitage suivant :

$$\tilde{f} = f + \xi,$$

où $\xi \sim \mathcal{N}(0, \sigma^2)$. Le but du débruitage est de construire un estimateur $\hat{f}$ de f qui ne dépend que des observations $\tilde{f}$.

Une façon de construire un estimateur non linéaire est de considérer une procédure de seuillage dans un espace transformé défini à l'aide d'une *frame*, c'est-à-dire une famille $\mathfrak{F} = \{r_i\}_{i \in I}$ de vecteurs de $\mathbb{R}^n$ telle qu'il existe $A, B > 0$ satisfaisant pour tous les $f \in \mathbb{R}^n$

$$A\|f\|_2^2 \leq \sum_{i \in I} |\langle f, r_i \rangle|^2 \leq B\|f\|_2^2.$$

Une *frame* est dite *ajustée* (ou *étroite*) si $A = B$.

La matrice $\mathcal{L}$ est symétrique, sa décomposition spectrale est donnée par $\mathcal{L} = \sum_\ell \lambda_\ell \langle \chi_\ell, \cdot \rangle \chi_\ell$, où $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = 0$ désignent les valeurs propres (ordonnées) de la matrice $\mathcal{L}$, et $(\chi_\ell)_{1 \leq \ell \leq n}$ sont les vecteurs propres associés. Ainsi, pour toute fonction $\rho : \mathrm{sp}(\mathcal{L}) \to \mathbb{R}$ définie sur le spectre de $\mathcal{L}$, on a $\rho(\mathcal{L}) = \sum_\ell \rho(\lambda_\ell) \langle \chi_\ell, \cdot \rangle \chi_\ell$.

Les valeurs propres λ_ℓ, $\ell = 1, \ldots, n$ s'interprètent comme les fréquences fondamentales du graphe et $\rho(\mathcal{L})$ comme un opérateur de filtrage en termes d'analyse du signal.

La *frame* ajustée considérée ici est construite à partir d'une partition de l'unité finie $(\psi_j)_{j=0,\ldots,J}$ du compact $[0, \lambda_1]$ définie comme suit : soit $\omega : \mathbb{R}^+ \to [0,1]$ une fonction quelconque à support dans $[0,1]$, satisfaisant $\omega \equiv 1$ sur $[0, b^{-1}]$, pour $b > 1$, et

$$\psi_0(x) = \omega(x), \quad \psi_j(x) = \omega(b^{-j}x) - \omega(b^{-j+1}x) \ \text{ pour } \ j = 1, \ldots, J, \ \text{ où } \ J = \left\lfloor \frac{\log \lambda_1}{\log b} \right\rfloor + 2.$$

Il est montré dans Göbel et al. (2018) que l'ensemble de vecteurs suivants est une *frame* ajustée :

$$\mathfrak{F} = \left\{ \sqrt{\psi_j}(\mathcal{L}) \delta_i, j = 0, \ldots, J, i = 1, \ldots, n \right\}.$$

La transformée en ondelettes d'un signal $f \in \mathbb{R}^n$ est donnée par

$$\mathcal{W}f = \left(\sqrt{\psi_0}(\mathcal{L}) f^T, \ldots, \sqrt{\psi_J}(\mathcal{L}) f^T \right)^T \in \mathbb{R}^{n(J+1)},$$

et sa transformée inverse $\mathcal{W}^*$ est donnée par

$$\mathcal{W}^* \left(\eta_0^T, \eta_1^T, \ldots, \eta_J^T \right)^T = \sum_{j \geq 0} \sqrt{\psi_j}(\mathcal{L}) \eta_j.$$

Étant donné le laplacien et une *frame* donnée, le débruitage dans ce contexte peut être résumé comme suit :

- Analyse : calculer la transformée $\mathcal{W}\tilde{f}$;
- Seuillage : appliquer un opérateur de seuillage donné (par ex. *dur* ou *doux*) aux coefficients $\mathcal{W}\tilde{f}$;
- Synthèse : appliquer la transformée inverse pour obtenir une estimation $\hat{f}$ de f.

Les performances dépendent fortement du choix du paramètre de seuillage. Le SURE a été étendue pour cette *frame* afin de sélectionner un seuil optimal en minimisant ce critère dans de Loynes et al. (2021). Le principal inconvénient de cette approche est qu'elle nécessite le calcul de la diagonalisation complète du laplacien si bien que son application se limite aux graphes de taille modérée. Les difficultés à lever se situent à la fois au moment de la transformation et lors de la phase d'optimisation du SURE dont le terme de divergence fait apparaître des poids caractéristiques de la transformée. La première difficulté a déjà été traitée dans Hammond et al. (2011) en proposant une transformée rapide à base d'approximations par polynômes de Chebyshev. Le deuxième écueil, quant à lui, sera résolu par Monte Carlo en profitant de l'interprétation en terme de covariance des poids. Ces points font l'objet de la section suivante.

3 SURE adapté aux grands graphes

Les polynômes de Chebyshev de première espèce $T_k(x)$ d'ordre k peuvent être calculés par la relation de récurrence $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, avec $T_0(x) = 1$ et $T_1(x) = x$. Ces polynômes forment une base orthogonale de l'espace de Hilbert $\mathbb{L}^2([-1,1], dy/\sqrt{1-y^2})$. Une ondelette (ou filtre) ρ peut donc être approchée par le développement tronqué d'ordre $K-1$,

$$\rho_K(\mathcal{L}) = \sum_{i=0}^{K-1} \theta_i(\rho) T_i(\mathcal{L}),$$

où $\theta_i(\rho)$ est le coefficient de Chebyshev associé à $T_i(\mathcal{L})$, le i-ème polynôme de Chebyshev évalué en $\tilde{\mathcal{L}} = 2\mathcal{L}/\lambda_1 - I_n$. En suivant Hammond et al. (2011), pour tout filtre ρ défini sur $\mathrm{sp}(\mathcal{L})$ et tout signal f sur le graphe $\mathcal{G}$, l'approximation $\rho_K(\mathcal{L})f$ fournie est proche de $\rho(\mathcal{L})f$ avec une complexité en temps $O(|\mathcal{E}|K)$ raisonnable lorsque la matrice $\mathcal{L}$ est creuse.

D'après de Loynes et al. (2021), le SURE pour un processus général de seuillage $h : \mathbb{R}^{n(J+1)} \to \mathbb{R}^{n(J+1)}$ est donné par l'identité suivante

$$\mathbf{SURE}(h) = -n\sigma^2 + \|h(\widetilde{F}) - \widetilde{F}\|^2 + 2\sum_{i,j=1}^{n(J+1)} \gamma_{i,j}^2 \partial_j h_i(\widetilde{F}), \tag{1}$$

où $\gamma_{i,j}^2 = (\mathcal{W}\mathcal{W}^*)_{i,j}$. Dans de Loynes et al. (2021), les poids $\gamma_{i,j}^2$, $i,j = 1,\ldots,n(J+1)$, sont calculés à partir de la réduction complète de la matrice laplacienne qui n'est plus réalisable

pour les grands graphes. En outre, il est clair d'après l'interprétation probabiliste donnée dans de Loynes et al. (2021) que

$$\forall i, j = 1, \ldots, n(J+1), \quad \gamma^2_{i,j} = \mathbb{E}[(\mathcal{W}\varepsilon)_i(\mathcal{W}\varepsilon)_j]$$

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ sont n variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*) de moyenne nulle et de variance égale à un. Ainsi, en tirant parti de cette identité, les poids peuvent être estimés comme suit

- Générer N variables aléatoires *i.i.d.* $(\varepsilon_{i,k})_{i=1,\ldots,n,k=1,\ldots,N}$ telles que $\mathbb{E}[\varepsilon_{i,k}] = 0$ et $\mathbb{V}(\varepsilon_{i,k}) = 1$;
- Calculer

$$\hat{\gamma}^2_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\left(\sum_{p=1}^{n}\mathcal{W}_{i,p}\varepsilon_{p,k}\right)\left(\sum_{p=1}^{n}\mathcal{W}_{j,p}\varepsilon_{p,k}\right).$$

4 Simulations

Nous comparons notre méthode à celle du *DFS Fused Lasso* introduite dans Padilla et al. (2017) et Rudin et al. (1992) par des simulations numériques réalisées avec les packages R **igraph** (Csardi and Nepusz 2006), **genlasso** (Arnold and Tibshirani 2020) et **gasper** (de Loynes et al. 2020). Ce dernier fournit une interface à la *SuiteSparse Matrix Collection*[1] (Davis and Hu 2011). Ici, nous utilisons le graphe des routes de l'état de Pennsylvanie[2] tiré de Leskovec et al. (2009) composé de 1088092 nœuds et de 1541898 arêtes. Sur celui-ci, deux classes de signaux synthétiques sont générées en s'inspirant de la méthodologie introduite dans Behjat et al. (2016). Selon deux paramètres $p \in (0, 1)$ et $k \in \mathbb{N}$, on produit un signal $f_{p,k} = W^k x_p/\lambda_1^k$ où W est la matrice de poids, x_p la réalisation d'une variable aléatoire suivant une loi de Bernoulli de paramètre p et λ_1 la plus grande valeur propre de $\mathcal{L}$. Pour ces simulations, nous avons généré deux signaux avec les paramètres $p = 0.001$, $k = 4$ et $p = 0.01$, $k = 10$ respectivement.

Table 1: SNR moyen calculé sur 10 réalisations pour chaque niveau de bruit.

	$f_{0.001,4}$				$f_{0.01,10}$			
SNR_{in}	0.61	6.63	12.65	18.67	2.09	8.12	14.13	20.16
MSE_{LD}	12.28	17.41	21.37	23.53	9.19	13.33	16.51	18.01
SURE_{LD}	12.28	17.41	21.37	23.53	9.18	13.33	16.51	18.01
MSE_{DFS}	7.92	13.10	18.41	24.27	6.64	11.12	15.98	18.46

Les performances sont comparées en termes de rapport signal sur bruit (SNR) pour différents niveaux de bruits et pour chaque signal synthétique $f_{p,k}$. L'analyse multi-échelle présente globalement des résultats très prometteurs en comparaison à *DFS Fused Lasso*.

[1] `https://sparse.tamu.edu/`

[2] `https://sparse.tamu.edu/SNAP/roadNet-PA`

Bibliographie

Arnold, T. B. and R. J. Tibshirani (2020). *R package genlasso: Path algorithm for generalized lasso problems*. Version 1.5.

Behjat, H., U. Richter, D. Van De Ville, and L. Sörnmo (2016). Signal-adapted tight frames on graphs. *IEEE Transactions on Signal Processing 64*(22), 6017–6029.

Bronstein, M. M., J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag. 34*(4), 18–42.

Coifman, R. R. and M. Maggioni (2006). Diffusion wavelets. *Applied and Computational Harmonic Analysis 21*(1), 53 – 94. Special Issue: Diffusion Maps and Wavelets.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Davis, T. A. and Y. Hu (2011). The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS) 38*(1), 1.

de Loynes, B., F. Navarro, and B. Olivier (2020). Gasper: Graph signal processing in r.

de Loynes, B., F. Navarro, and B. Olivier (2021). Data-driven thresholding in denoising with Spectral Graph Wavelet Transform. *J. Comput. Appl. Math. 389*, 113319.

Dong, X., D. Thanou, L. Toni, M. Bronstein, and P. Frossard (2020). Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Process. Mag. 37*(6), 117–127.

Göbel, F., G. Blanchard, and U. von Luxburg (2018). Construction of tight frames on graphs and application to denoising. In *Handbook of big data analytics*, Springer Handb. Comput. Stat., pp. 503–522. Springer, Cham.

Hammond, D. K., P. Vandergheynst, and R. Gribonval (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis 30*(2), 129–150.

Leskovec, J., K. J. Lang, A. Dasgupta, and M. W. Mahoney (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics 6*(1), 29–123.

Nickel, M., K. Murphy, V. Tresp, and E. Gabrilovich (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE 104*(1), 11–33.

Ortega, A., P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE 106*(5), 808–828.

Padilla, O. H. M., J. Sharpnack, J. G. Scott, and R. J. Tibshirani (2017). The dfs fused lasso: Linear-time denoising over general graphs. *J. Mach. Learn. Res. 18*, 176–1.

Rudin, L. I., S. Osher, and E. Fatemi (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena 60*(1), 259–268.

Shuman, D., S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag. 3*(30), 83–98.

Modèle de régression par spline monotone pour des données de protéomique quantitative

Marie Chion [1], Joanna Bons [2], Myriam Maumy-Bertrand [3], Christine Carapito [4] & Frédéric Bertrand [5]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501, CNRS - Université de Strasbourg et Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, CNRS - Université de Strasbourg, Strasbourg, France. chion@math.unistra.fr*

[2] *Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, UMR 7178, CNRS – Université de Strasbourg, Strasbourg, France et The Buck Institute for Research on Aging, Novato, California, USA. JBons@buckinstitute.org*

[3] *Laboratoire de Modélisation et Sûreté des Systèmes, Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France. myriam.maumy@utt.fr*

[4] *Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, UMR 7178, CNRS – Université de Strasbourg, Strasbourg, France. ccarapito@unistra.fr*

[5] *Laboratoire de Modélisation et Sûreté des Systèmes, Institut Charles Delaunay, Université de Technologie de Troyes, Troyes, France. frederic.bertrand@utt.fr*

Résumé. L'analyse protéomique consiste à étudier les protéines d'un système biologique donné, à un moment donné et dans des conditions données. Les méthodes de quantification globale permettent de comparer des milliers de niveaux d'expression de protéines dans les différents échantillons biologiques considérés. Les méthodes de quantification ciblée permettent, par l'introduction de standards synthétiques marqués correspondant à des peptides d'intérêt préalablement sélectionnés, de connaître précisément la quantité de certaines protéines dans l'échantillon biologique considéré. Une approche récente, appelée Data-Independent Acquisition, permet de combiner ces deux méthodes en une seule analyse. En protéomique quantitative, l'hypothèse forte est faite d'une relation de proportionnalité entre la quantité d'un peptide et son intensité par le biais du facteur de réponse, spécifique à chaque peptide dans chaque échantillon. À partir des données d'intensité et de quantité obtenues en quantification ciblée, nous proposons d'ajuster des modèles de lissage par spline monotone expliquant la quantité d'un peptide par son intensité dans l'échantillon considéré. Ces modèles nous permettent ensuite d'estimer les quantités de tous les peptides dont l'intensité a été mesurée lors de l'étape de quantification globale.

Mots-clés. Régression, Spline monotone, Protéomique, Spectrométrie de masse quantitative.

Abstract. Proteomic analysis consists in studying proteins from a given biological system, at a given time and under given conditions. Global quantification methods make

it possible to compare thousands of proteins expression levels across the different biological samples that are considered. Targeted quantification methods allow, by introducing labelled synthetic standards corresponding to previously selected peptides of interest, to know precisely the quantity of specific proteins in the biological sample considered. A recent approach, called Data-Independent Acquisition, enables to combine these two methods in a single analysis. In quantitative proteomics, the strong hypothesis is made of a relationship of proportionality between the quantity of a peptide and its intensity through the response factor, specific to each peptide in each sample. From the intensity and quantity data obtained in targeted quantification, we propose to fit monotone spline models explaining the quantity of a peptide by its intensity in the considered sample. These models then allow us to estimate the amounts of all peptides that were detected in the global quantification step.

Keywords. Regression, Monotone spline smoothing, Proteomics, Quantitative mass spectrometry.

1 Expérience

Le travail décrit ici s'appuie sur l'expérience de Bonnet *et al.* (2020). 64 échantillons de muscles bovins pour lesquels 20 peptides, correspondant aux 10 protéines candidats biomarqueurs pour la tendreté et le persillage de la viande bovine, ont été analysés par une approche DIA SWATH-MS (Ludwig *et al.*, 2018).

1.1 Quantification ciblée

Une première étape de quantification ciblée a permis, à partir de l'intensité mesurée par chromatographie liquide couplée à la spectrométrie de masse, de déterminer précisément la quantité des 20 peptides d'intérêt dans chacun des 64 échantillons considérés. Pour ce faire, la relation suivante a été utilisée.

$$Quantité\ du\ peptide = \frac{Quantité\ du\ peptide\ synthétique \times Intensité\ du\ peptide}{Intensité\ du\ peptide\ synthétique}$$

1.2 Quantification globale

Une seconde étape de quantification globale a permis de mesurer l'intensité de près de 5500 peptides dans les 64 échantillons considérés. En protéomique quantitative, une hypothèse forte est faite, selon laquelle la quantité d'un peptide est proportionnelle à son

intensité au travers d'un facteur de réponse. Celui-ci est spécifique au peptide considéré et à l'échantillon considéré. Ainsi, nous pouvons écrire :

$$Quantité\ du\ peptide = Intensité\ du\ peptide \times Facteur\ de\ réponse$$

2 Lissage par spline monotone

La méthode de lissage par spline monotone combine la régression par I-spline (Ramsay, 1988) avec l'estimation des paramètres par la méthode des moindres carrés non négatifs. Dans notre travail, les modèles utilisés sont des combinaisons linéaires de I-splines, tels que :

$$f(x) = \sum_i a_i I_i(x|k,t)$$

où les a_i sont les paramètres à estimer et les I_i constituent une base de fonctions I-splines.

Une fonction I-spline s'écrit comme l'intégrale d'une fonction M-spline (fonction non-négative polynomiale par morceaux) :

$$I_i(x|k,t) = \int_L^x M(u|k,t)\,du$$

où k est le degré de la I-spline et L est la borne inférieure du domaine.

L'algorithme d'estimation des paramètres utilisé est celui de Lawson-Hanson (1995) pour les moindres carrés non négatifs, implémenté dans la fonction `nnls` du package "nnls" (Mullen et Van Stokkum, 2012) du logiciel libre `R`, version 4.0.2. L'analyse par I-spline a été réalisée à l'aide de la fonction `iSpline` du package "splines2", développé par Wang et Yan (2020).

3 Résultats

Un modèle de régression I-spline a été ajusté pour chacun des 64 échantillons bovins considérés, en utilisant les données obtenues à l'étape de quantification ciblée. Un exemple de ces ajustements est donné à la figure 1. Le logarithme de l'intensité des peptides a été choisie comme variable explicative et le logarithme de la quantité des peptides comme variable réponse.

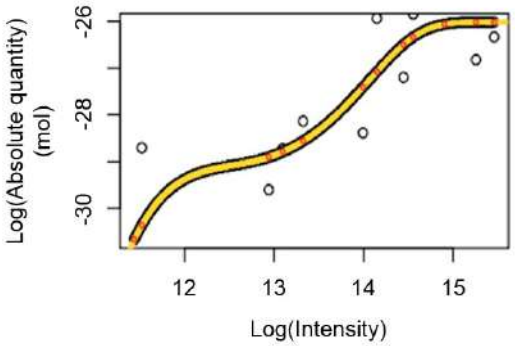


Figure 1: Exemple pour un des 64 échantillons du lissage par spline monotone et de la prédiction du logarithme de la quantité de peptides en fonction du logarithme de l'intensité des peptides. Les valeurs décrites par des cercles rouges correspondent valeurs prédites pour les données observées, qui sont représentées en noir.

Comme l'illustre l'exemple de la figure 1, l'hypothèse d'une relation linéaire entre l'intensité et la quantité ne peut être retenue. Sur les 64 échantillons considérés, la régression spline monotone offre un ajustement satisfaisant. En outre, elle surpasse la régression linéaire usuelle en termes d'erreur quadratique moyenne, comme le montre la figure 2.

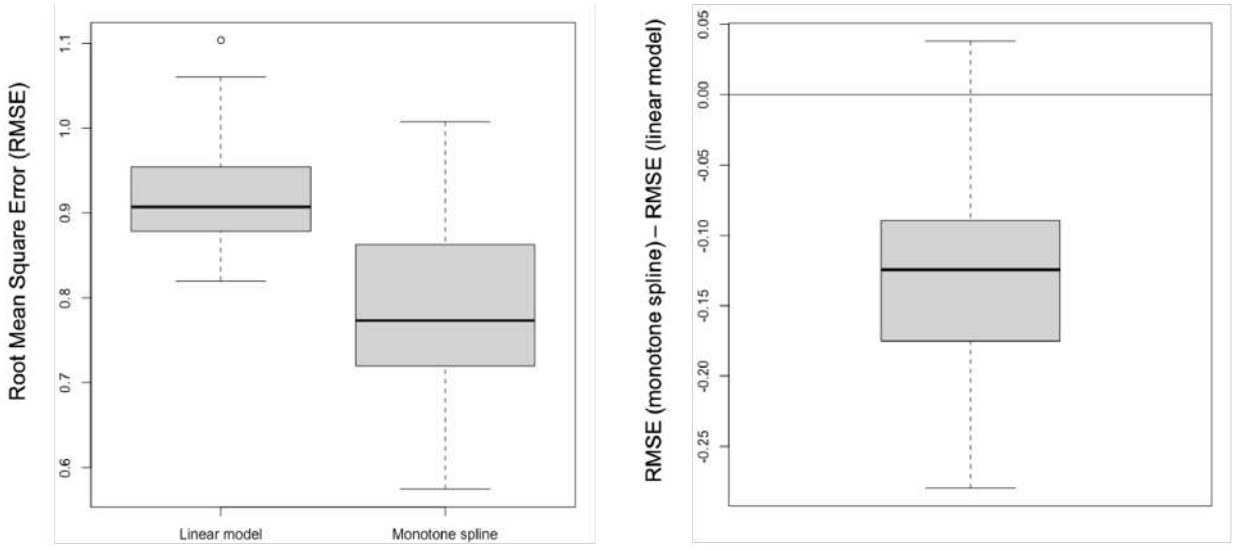


Figure 2: Comparaison des performances en termes d'erreur quadratique moyenne du modèle de spline monotone par rapport au modèle de régression linéaire simple.

Ensuite, à partir des intensités des protéines quantifiées lors de l'étape de quantification ciblée, nous avons estimé leur quantité absolue. Les prédictions sont considérées comme étant de bonne qualité si le rapport quantité estimée/quantité réelle est compris entre 0,5 et 2. La vérification des quantités prédites sur les 20 peptides mesurés en quantification ciblée est représentée sur la figure 3. Ainsi, 53% des rapports quantité prédite/quantité réelle sont compris entre 0,5 et 2, illustrés par la bande rouge.

Une étape supplémentaire d'analyse biologique des quantités absolues prédites a montré que nos résultats étaient conformes à la littérature sur le protéome des muscles bovins (Bons *et al.*, 2021).

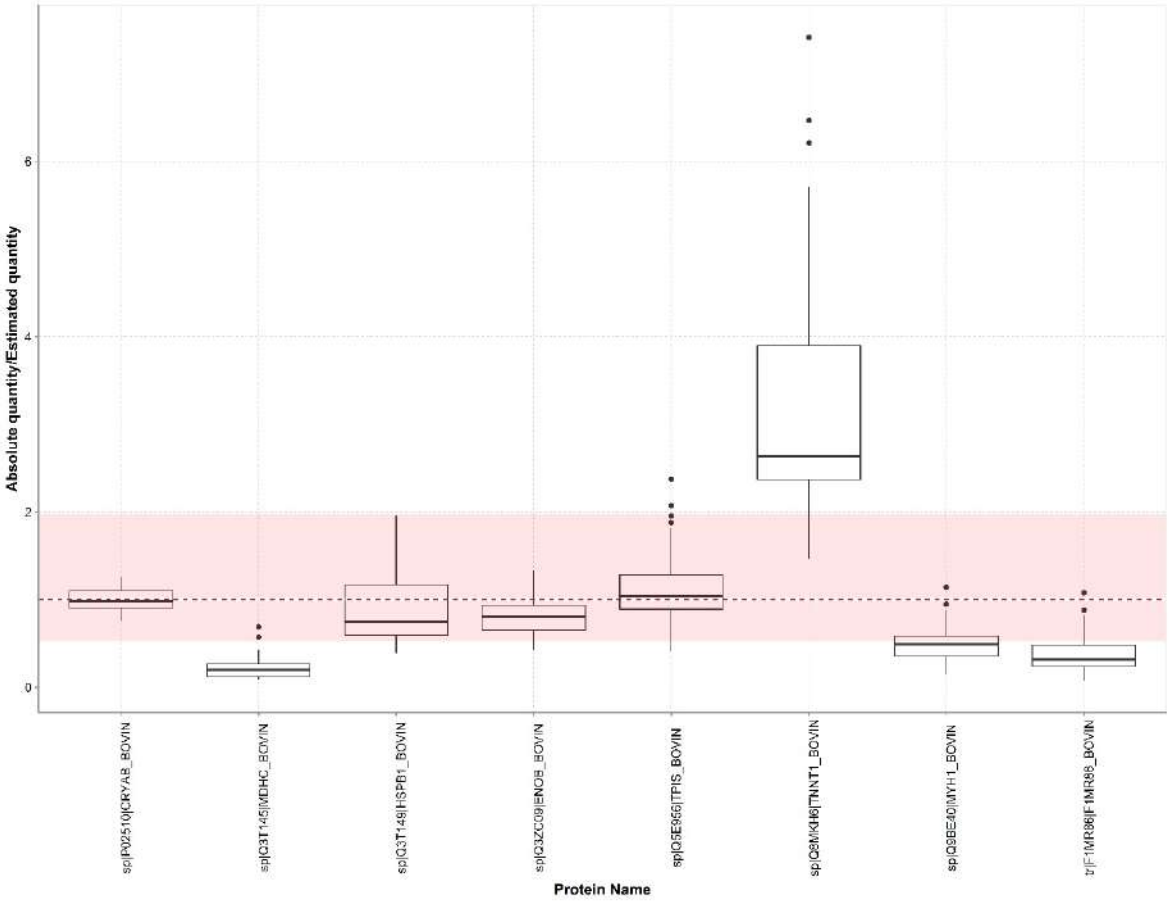


Figure 3: Diagrammes en boîte du rapport de la quantité prédite sur la quantité réelle pour chaque protéine d'intérêt dans chacun des 64 échantillons considérés.

Bibliographie

Bonnet, M., Soulat, J., Bons, J., Léger, S., De Koning, L., Carapito, C. & Piccard, B. Quantification of biomarkers for beef meat qualities using a combination of Parallel Reaction Monitoring- and antibody-based proteomics. *Food Chemistry*, 317, 2020.

Bons, J, Husson, G, Chion, M, Bonnet, M., Maumy-Bertrand, M., Delalande, F., Cianférani, S., Bertrand, F., Picard, B. & Carapito, C.. (2021). Combining label-free and label-based accurate quantifications with SWATH-MS: Comparison with SRM and PRM for the evaluation of bovine muscle type effects. *Proteomics*. e2000214.

Lawson, C. L. & Hanson, R. J. (1995). *Solving Least Squares Problems*. SIAM.

Ludwig, C., L. Gillet, L., Rosenberger, G., Amon, S., Collins, B.C. & Aebersold, R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 14(8), 2018.

Mullen, K. M. & Van Stokkum, I. H. M. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4. 2012.

Ramsay, J.O. Monotone Regression Splines in Action, *Statistical Science*, 3(4), 425-461, 1988.

Wang, W. & Yan, J.: splines2: Regression Spline Functions and Classes. R package version 0.3.1. 2020.

Analyse de sensibilité lors de la généralisation d'un effet causal

Bénédicte Colnet [1], Julie Josse[2], Erwan Scornet [3] & Gaël Varoquaux[1]

[1] *Université Paris-Saclay, Inria - benedicte.colnet@inria.fr, gael.varoquaux@inria.fr*
[2] *Inria, Sophia-Antipolis - julie.josse@inria.fr*
[3] *CMAP, Ecole Polytechnique - erwan.scornet@polytechnique.edu*

Résumé. Lorsque l'on souhaite généraliser l'effet moyen de traitement estimé lors d'un essai randomisé contrôlé à une population cible d'intérêt, plusieurs stratégies existent. Elles consistent soit à pondérer la population de l'essai randomisé de façon à ressembler à la population cible (IPSW), soit à modéliser la fonction liant la variable d'intérêt et les covariables (g-formula). La pratique suppose de réunir au moins deux jeux de données différents - un essai randomisé et une étude observationnelle - ce qui peut impliquer un faible nombre de variables communes. Or ceci implique une potentielle perte d'identifiabilité de l'effet de traitement sur la population cible. Nous proposons ici de traiter les différents cas pratiques où une covariable majeure est manquante dans les deux jeux de données, ou alors partiellement manquante. Pour cela nous adaptons les analyses de sensibilité plus classiques de l'inférence causale au cas singulier de la généralisation d'un effet de traitement. Nous détaillons le cas linéaire mais aussi semi-paramétrique. Les approches sont illustrées par des simulations et motivées par des données réelles.

Mots-clés. Effet de traitement moyen (ATE), validité interne, validité externe, variable manquante, essai randomisé contrôlé (RCT), inférence causale.

Abstract. When generalizing an average treatment effect measured on a randomized controlled trial (RCT) to a target population of interest, several off-the-shell methods exist. A first method proposes to reweight the RCT sample so that it resembles the target population with respect to the observed covariates and plausible moderators (IPSW). A second method proposes to model the outcome with and without treatment conditionally to the covariates, and then to extend the model to the target population of interest (g-formula). But in practice, once the two datasets of interest to answer the question are gathered - a RCT and an observational study -, the number of common variables measured between the observational data and the RCT may be small. Taking the subset of common covariates may break key assumptions necessary to generalize the treatment effect estimated. We propose to address the different practical cases when a covariate is either missing in both or one data set. The linear and semi-parametric cases are detailed. Simulations along with real data analysis are provided.

Keywords. Average treatment effect (ATE), external validity, internal validity, unobserved covariate, randomized controlled trial (RCT), causal inference.

1 Contexte

1.1 Motivation

Les politiques publiques s'appuient fortement sur les essais randomisés contrôlés pour évaluer les bénéfices d'une certaine politique, car ils permettent de répondre à des questions telles que *devrait-on ou non administrer ce médicament à la population* ? Ainsi, les essais randomisés sont aujourd'hui la méthodes de référence pour conclure sur la causalité d'une certaine variable sur un résultat (Imbens and Rubin, 2015). Cependant, ils sont parfois irréalisable, ou contraire à l'éthique. De plus, même lorsque des essais randomisés sont conduits, ils peuvent souffrir de critères d'éligibilité trop stricts, ce qui rend la population sélectionnée sensiblement différente de la population cible sur laquelle le législateur entend appliquer la politique. C'est pourquoi des méthodes permettant d'exploiter à la fois les informations des essais randomisés et des données observationnelles ont été proposées afin de *généraliser* l'effet moyen de traitement (Cole and Stuart, 2010; Stuart et al., 2011; Bareinboim and Pearl, 2013; Tipton, 2013; Bareinboim and Pearl, 2016; Kallus et al., 2018; Dong et al., 2020; Cinelli and Pearl, 2020). En pratique cela suppose d'avoir au moins deux jeux de données, typiquement un essai randomisé pour le traitement d'intérêt, ainsi qu'une étude observationnelle représentative de la population d'intérêt.
Mais en pratique, une fois que les jeux de données sont réunis pour répondre à la question, le nombre de variables mesurées de manière consistente peut être faible. Une pratique courante consiste à prendre le sous-ensemble des covariables communes pour généraliser, ce qui peut briser les hypothèses d'identifiabilité. Dans ce cadre, l'une des solutions est de s'inspirer des analyses en sensibilité, afin d'évaluer l'impact d'un facteur confondant non observé (Imbens, 2003; Rosenbaum, 2005; Franks et al., 2019; Veitch and Zaveri, 2020). À notre connaissance, cela n'a pas, ou peu, été développé dans le cas de la généralisation d'un effet de traitement d'un essai randomisé vers une population cible, sauf par Nguyen et al. (2018).

1.2 Contributions

Nous étendons les approches d'analyse de sensibilité développées dans le cadre de l'analyse d'un jeu de données observationnel, en prenant en compte la spécificité de la question par rapport au champ d'où sont issues ces méthodes. En particulier les critères de sensibilité utilisés ont une interprétation différente, tout comme les hypothèses initiales. Par ailleurs, nous prenons aussi en compte le fait que - contrairement aux analyses de sensibilité visant un confondant complètement caché - la variable manquante est parfois partiellement manquante et peut-être utilisée pour mieux interpréter et préciser l'analyse en sensibilité. À notre connaissance aucune analyse de sensibilité n'a été proposée pour ce problème, en dehors de la méthode proposée par Nguyen et al. (2018) qui concerne le cas linéaire pour

une variable manquante dans le jeu de données observationnel seulement. En particulier, nous obtenons des résultats quantitatifs dans le cadre linéaire où les conditions de biais et l'ampleur du biais sont explicités comme illustré dans ce document.

1.3 Formalisation du problème

Le problème est formalisé dans le cadre général de Neyman-Rubin, où nous modélisons chaque patient dans l'essai randomisé ou la population d'observation comme décrit par un 5-uplet

$$(X, Y(0), Y(1), A, S)$$

tiré d'une distribution $(X, Y(0), Y(1), A, S) \in \mathcal{X} \times \mathcal{Y}^2 \times \{0,1\} \times \{0,1\}$.

X est un vecteur de taille p, A correspond au traitement (que nous considérons ici binaire avec $A = 0$ pour l'absence de traitement et $A = 1$ pour le traitement), $Y(a)$ est la variable continuous d'intérêt (*outcome*) pour un certain traitement a (pour $a \in \{0,1\}$), et S est l'indicatrice de l'éligibilité dans l'essai randomisé. On définit l'effet moyen conditionnel du traitement (CATE) par :

$$\forall x \in \mathcal{X}\,, \quad \tau(x) = \mathbb{E}[(1) - Y(0) \mid X = x] \tag{1}$$

et l'effet moyen du traitement de la population cible par:

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\tau(X)] \tag{2}$$

Comme les observations de l'essai et les données d'observation ne suivent pas la même distribution, τ est différent de l'effet de traitement moyen de l'essai randomisé τ_1 qui peut être exprimé mathématiquement comme suit :

$$\tau \neq \tau_1 = \mathbb{E}[Y(1) - Y(0) \mid S = 1]\,.$$

Dans le cas que nous traitons, les variables observées dans chacun des jeux de données sont différentes. Ainsi on peut décomposer X tel que $X = X_m \cup X_o$, où X_o est le set de variables observées dans les deux jeux de données, et X_m est le set de variables manquantes ou partiellement manquantes dans la réunion des deux jeux de données.

Une hypothèse majeure dans le cas de la généralisation avec toutes les données est l'hypothèse dite d'ignorabilité.

$$\{Y(0), Y(1)\} \perp S \mid X_o, X_m$$

Or, ici si on considère que le set de variables n'est pas complet, cette hypothèse peut ne plus être valide, soit:

$$\{Y(0), Y(1)\} \not\perp S \mid X_o$$

La Figure 1 présente la structure type des données traitées.

	Covariables				
	X_1	X_2	X_3	A	Y
1	1.1	20	5.4	1	10
	-6	45	8.3	0	12.4
n	0	15	6.2	1	18
$n+1$		...		...	...
	-2	52	NA	NA	NA
	-1	35	NA	NA	NA
$n+m$	-2	22	NA	NA	NA

	Covariables				
	X_1	X_2	X_3	A	Y
1	1.1	20	NA	1	32
	-6	45	NA	0	1.3
n	0	15	NA	1	63
$n+1$		...		...	...
	-2	52	3.4	NA	NA
	-1	35	3.1	NA	NA
$n+m$	-2	22	5.7	NA	NA

Figure 1: Structure de données type considérée lorsqu'une covariable supplémentaire - quantitative ou catégorielle - serait disponible dans l'essai randomisé, mais pas dans l'ensemble de données observationnel (à gauche) ou dans la situation inverse (à droite).

1.4 Estimateurs

Plusieurs estimateurs permettent de généraliser l'effet moyen de traitement (Colnet et al., 2020), et reposent notamment sur une hypothèse dite de transportabilité.

Hypothèse 1.1 (Transportabilité du CATE)**.** *Pour tout* $x \in \mathcal{X}, \mathbb{E}[Y(1) - Y(0) \mid X = x, S = 1] = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau(x)$

C'est cette même hypothèse qui est brisée lorsque des variables sont manquantes. Parmis tous les estimateurs possibles, nous pouvons en présenter la *g-formula* qui consiste à modéliser le CATE (Equation 1) en utilisant le jeu de données expérimental (RCT), puis d'intégrer cette formule sur le jeu de données observationnelles.

$$\widehat{\tau}^g = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\widehat{\tau}(X_i)\right), \tag{3}$$

où $\widehat{\tau}(X)$ est un estimateur de $\tau(X)$ obtenu sur le RCT.

2 Extrait: contribution dans le cas linéaire

Résultats théoriques Supposons que Y dépende linéairement de X, c'est-à-dire qu'il existe $\alpha, \beta \in \mathbb{R}^p$, et $\sigma \in \mathbb{R}^+$ tels que:

$$Y = \langle X, \beta \rangle + A\langle X, \alpha \rangle + \varepsilon, \text{ où } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \tag{4}$$

Nous ajoutons une hypothèse de distribution multivariée gaussienne de X, c'est-à-dire que, $X \sim \mathcal{N}(\mu, \Sigma)$, ainsi qu'une hypothèse de transportabilité de la matrice de variance-covariance que nous notons Σ (c'est-à-dire que nous supposons que cette matrice

de variance covariance est la même dans les deux jeux de données). En particulier, la matrice $\Sigma_{m,o}$ correspond à la sous-matrice de Σ avec les lignes correspondants aux variables manquantes m et les colonnes aux variables observées o.

Dans ce cas, le biais de l'estimateur g-formula (équation 3), mais appliqué uniquement sur les covariables complètement observées, que l'on note τ_o^g, peut être calculé explicitement:

$$\tau - \tau_o^g = \sum_{j \in manquant} \alpha_j \left(\mathbb{E}[X_m] - \mathbb{E}[X_m \mid S = 1] - \Sigma_{m,o}\Sigma_{o,o}^{-1}(\mathbb{E}[X_o] - \mathbb{E}[X_o \mid S = 1]) \right) \quad (5)$$

Simulations Ce résultat est illustrable par des simulations. Soient $X_j \sim \mathcal{N}(1, 1)$ pour tout $j = 1, \ldots, p$. On modélise la sélection dans l'essai randomisé de la façon suivante:

$$\text{logit}\left\{\mathbb{P}[S = 1 \mid X]\right\} = \beta_{s,0} + \beta_{s,1} X_1 + \cdots + \beta_{s,p} X_p.$$

Et la variable Y suit le modèle génératif suivant:

$$Y(a) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + a(\alpha_1 X_1 + \cdots + \alpha_p X_p) + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0, 1).$$

Les paramètres utilisés sont $p = 5$, $\beta = (5, 5, 5, 5, 5)$, et plus amplement détaillés dans le tableau 1.

Covariables	X_1	X_2	X_3	X_4	X_5
α	$\alpha_1 = 30$	$\alpha_2 = 30$	$\alpha_3 = 10$	$\alpha_4 = 0$	$\alpha_5 = 0$
β_s	$\beta_{s,1} = -0.3$	$\beta_{s,2} = 0$	$\beta_{s,3} = -0.3$	$\beta_{s,4} = -0.3$	$\beta_{s,5} = 0$
$. \perp X_1$	-	$X_2 \perp X_1$	$X_3 \perp X_1$	$X_4 \perp X_1$	$X_5 \not\perp X_1$

Table 1: Paramètres utilisés lors de la simulation d'un cas linéaire.

Dans ce cas, les biais observés dans les simulations correspondent aux biais théoriques attendus de l'équation 5. Les résultats des simulations sont présentés sur la Figure 2.

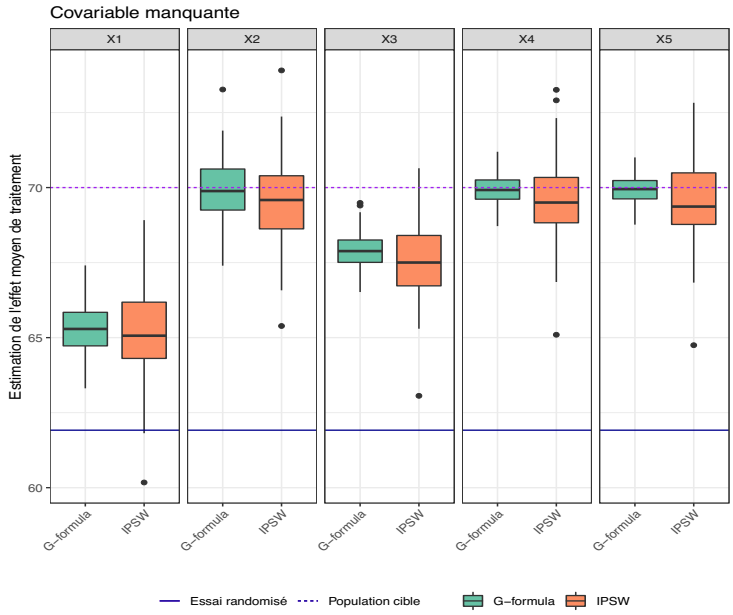


Figure 2: Biais des estimateurs IPSW et g-formula dans le cas de l'omission d'une variable sur la généralisation de l'effet de traitement. Les paramètres de la simulation sont présentés tableau 1, et pour 100 répétitions.

References

Bareinboim, E. and J. Pearl (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference 1*(1), 107–134.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*(27), 7345–7352.

Cinelli, C. and J. Pearl (2020, October). Generalizing experimental results by leveraging knowledge of mechanisms. *European journal of epidemiology*.

Cole, S. R. and E. A. Stuart (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American Journal of Epidemiology 172*, 107–115.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang (2020). Causal inference methods for combining randomized trials and observational studies: a review.

Dong, L., S. Yang, X. Wang, D. Zeng, and J. Cai (2020). Integrative analysis of randomized clinical trials with real world evidence studies. *arXiv preprint arXiv:2003.01242*.

Franks, A., A. D'Amour, and A. Feller (2019, 04). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 1–38.

Imbens, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge UK: Cambridge University Press.

Kallus, N., A. M. Puli, and U. Shalit (2018). Removing hidden confounding by experimental grounding. In *Advances in neural information processing systems*, pp. 10888–10897.

Nguyen, T., B. Ackerman, I. Schmid, S. Cole, and E. Stuart (2018, 12). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLOS ONE 13*, e0208795.

Rosenbaum, P. (2005, 10). *Sensitivity Analysis in Observational Studies*, Volume 4.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*, 369–386.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics 38*, 239–266.

Veitch, V. and A. Zaveri (2020). Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding.

Une approche permettant de maîtriser le niveau de confiance en optimisation multi-objectifs "data-driven"

Alexandre Conanec [1,2], Marie Chavent [2], Marie-Pierre Ellies-Oury [1] & Jérôme Saracco [2]

[1] *INRAE Biomarqueur team, Theix, 63122 Saint Genès Champanelle, France & Bordeaux Sciences Agro, 33175 Gradignan, France.*
alexandre.conanec@agro-bordeaux.fr ; marie-pierre.ellies@agro-bordeaux.fr
[2] *Inria BSO, ASTRAL team & IMB 5251 CNRS, Université de Bordeaux & Bordeaux INP, 33400 Talence, France.*
marie.chavent@math.u-bordeaux.fr ; jerome.saracco@inria.fr

Résumé. L'optimisation dite "data-driven" permet de résoudre des problèmes pour lesquels le lien entre la variable de décision et l'objectif est inconnu en utilisant un modèle dit de substitution. Ce modèle est estimé à partir d'observations disponibles et est généralement entaché d'une erreur. Une gestion intuitive de cette incertitude est la "value-at-risk" assurant que la valeur optimale est espérée (au sens de l'espérance mathématique) dans $1 - \tau\%$ des cas. Dans un contexte d'optimisation multi-objectifs, nous présentons une approche permettant d'ajuster le niveau τ du quantile sous-jacent de chaque objectif afin de maîtriser le niveau de confiance global.

Mots-clés. Optimisation, Multi-objectifs, Modèle de substitution, Régression quantile, "Value-at-risk"

Abstract. Data-driven optimization solves problems for which the link between the decision variable and the objective is unknown, by using a surrogate model. This model is estimated from observations and is generally noised. An intuitive way of dealing with this uncertainty is the value-at-risk ensuring that the optimal value is expected in $1 - \tau$ % of cases. In a context of multi-objective optimization, we present an approach to adjust the risk τ of the underlying quantile of each objective in order to control the overall confidence level.

Keywords. Optimization, Multi-objectives, Surrogate model, Quantile regression, Value-at-risk

Introduction

La complexité des mécanismes biologiques rend impossible la formulation analytique de certains problèmes d'optimisation en agriculture (Conanec et al. 2020). Une approche raisonnable consiste à considérer le mécanisme sous forme de "boîte noire" et d'en approcher les variations avec un modèle de substitution m (McBride et al. 2019) à partir d'observations du couple $(y, \mathbf{x})$

$$y = m(\mathbf{x}) + \epsilon, \tag{1}$$

où y est la variable réelle représentant l'objectif, $\mathbf{x}$ est la variable d-dimensionnelle désignant les prédicteurs, et ϵ est un terme d'erreur aléatoire matérialisant l'insuffisance des prédicteurs $\mathbf{x}$ à expliquer les variations de y dans le cadre de ce modèle. Parmi les différentes stratégies pour gérer cette incertitude, la "Value-at-Risk" (VaR, qui n'a pas vraiment une traduction en français), mobilisant un quantile associé à un risque τ, permet une interprétation intuitive de l'optimum : ce dernier est atteint dans $1-\tau$ % des cas (Bertsimas et al. 2006).
Par ailleurs, le décideur souhaite souvent optimiser plusieurs fonctions objectifs qui peuvent être antagonistes comme les performances de production et la qualité du produit. La gestion de l'incertitude dans ce problème d'optimisation multi-objectifs par la VaR est donc plus complexe. L'approche présentée dans cette communication vise à proposer une adaptation multi-objectifs de la VaR afin de garantir, dans les limites du possible, que le compromis optimal soit atteint dans $1-\tau$ % des cas.

1 Optimiser la VaR

L'optimisation à partir de données, dite optimisation "data-driven", nécessite la modélisation de la fonction objectif par un modèle de substitution estimé à partir d'un échantillon d'observations du couple $(y, \mathbf{x})$. Peu importe la qualité du modèle estimé, une erreur non réductible subsiste dans la prédiction de l'objectif. La VaR utilise la notion de quantile

$$\max_{\mathbf{x}} \; q_\tau(\mathbf{x}) = \max_{\mathbf{x}} \; \inf\{y | P(Y \leq y | X = \mathbf{x}) = \tau\} \tag{2}$$

garantissant que le maximum sera obtenu dans $1-\tau$ % des cas.
Un modèle de régression linéaire quantile, régularisé par la norme L_2, a été choisi par simplicité

$$q_\tau(\mathbf{x}) = \beta_n' \mathbf{x} = \sum_{k=1}^{d} \beta_{k,n} x_k, \text{ avec } \beta_n = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\beta) + \lambda \sum_{k=1}^{d} \beta_k^2, \tag{3}$$

où $\rho_\tau(u) = \tau u \mathbb{1}_{[u \geq 0]} - (1-\tau) u \mathbb{1}_{[u<0]}$ est la fonction de perte et λ est l'hyper-paramètre de régularisation qui sera calibré par validation croisée.

2 Adaptation à un contexte multi-objectifs

Un problème d'optimisation est dit multi-objectifs lorsque p objectifs sont à maximiser

$$\max_{\mathbf{x} \in \mathcal{X}} \quad \mathbf{q}_\tau(\mathbf{x}) = \max_{\mathbf{x} \in \mathcal{X}} \quad (q_\tau^{(1)}(\mathbf{x}), \ldots, q_\tau^{(p)}(\mathbf{x})). \tag{4}$$

N'existant pas d'ordre canonique dans $\mathbb{R}^p$, il est nécessaire de définir une relation de dominance pour tester la supériorité d'une solution par rapport à une autre, la plus

utilisée étant la dominance de Pareto (Ehrgott 2005). On dira que la solution $\mathbf{x}'$ domine $\mathbf{x}''$, notée $\mathbf{x}' \succ \mathbf{x}''$ si et seulement si

$$\forall\ j \in \{1,\ldots,p\},\ m_j(\mathbf{x}') \geq m_j(\mathbf{x}'') \quad \text{et} \quad \exists\ j \in \{1,\ldots,p\},\ m_j(\mathbf{x}') > m_j(\mathbf{x}''). \tag{5}$$

À partir de cette définition de la relation de dominance, on peut introduire la notion d'optimalité de Pareto comme l'ensemble

$$\mathcal{P}^* = \{\mathbf{x}^* \in \mathcal{X} \mid \nexists\ \mathbf{x} \in \mathcal{X} : \mathbf{x} \succ \mathbf{x}^*\} \tag{6}$$

et le front de Pareto par

$$\mathcal{PF}^* = \{\mathbf{q}_\tau(\mathbf{x}^*) \mid \mathbf{x}^* \in \mathcal{P}^*\}. \tag{7}$$

À l'optimalité, l'ensemble des solutions non dominées est retenu. Ce dernier peut être interprété comme l'ensemble des compromis possibles, autrement dit, aucune des solutions de cet ensemble n'est meilleure que les autres solutions sur tous les p objectifs. Le décideur peut alors choisir au sein de cet ensemble le compromis qui correspond le mieux à ses préférences. Pour trouver cet ensemble $\mathcal{P}^*$, le "Non-dominated Sorting Genetic Algorithm II" (NSGAII), classiquement utilisé pour résoudre ce type de problèmes multi-objectifs (Deb et al. 2000), a été mobilisé.

La gestion de l'incertitude par notre approche fondée sur des quantiles est perturbée par l'optimisation de plusieurs objectifs. En effet, la résolution du problème (4) permet juste d'affirmer que la valeur optimale de chaque objectif d'une solution est atteint dans $1-\tau$ % des cas :

$$P\left(y^{(j)*} \geq q_r^{(j)}(\mathbf{x}^*)\right) = 1-\tau,\ \forall j \in 1,\ldots,p, \tag{8}$$

où $y^{(j)*}$ est la valeur de l'objectif j pour une variable de décision fixée $\mathbf{x}^*$ qui est soumise à la réalisation de $\epsilon^{(j)}$, voir le modèle (1). Cependant, le décideur pourrait surtout vouloir garantir un niveau minimal de confiance pour l'ensemble des p objectifs avec une probabilité de réalisation maîtrisée

$$\Phi = P\left(\bigcap_{j=1}^{p} y^{(j)*} \geq q_r^{(j)}(\mathbf{x}^*)\right). \tag{9}$$

Notons que, comme illustré à la Figure 1, les relations entre les erreurs aléatoires $\epsilon^{(j)}$ font varier le niveau de confiance globale Φ. Par exemple, dans le cas d'indépendance entre les erreurs aléatoires $\epsilon^{(j)}$, on a $\Phi = (1-\tau)^p$.

3 Approche proposée

Pour pallier cette difficulté, nous proposons une approche permettant d'ajuster de manière itérative la valeur du risque τ en ajoutant une pénalité $\lambda^{(j)}$ pour chacun des p objectifs jusqu'à atteindre un niveau de confiance global de $1-\tau$. La procédure d'ajustement est détaillée dans l'algorithme 1.

Algorithm 1 Ajustement multi-objectifs du niveau de confiance

Input

X : matrice des variables explicatives de dimension $n \times d$

Y : matrice des variables à expliquer (objectifs) de dimension $n \times p$

τ : le niveau de risque compris tel que $0 < \tau < 1$

γ : un plancher en-dessous duquel $\tau + \lambda^{(j)}$ ne peut pas descendre

C, k : les paramètres (réels positifs) du pas $\eta(t) = Ce^{-kt}$

Output

$q^{(j)}_{\tau+\lambda_t^{(j)}}$: les p modèles de régression quantile associés à $\tau + \lambda_t^{(j)}$

$\tau + \lambda_t^{(j)}$: la valeur du risque pour chaque quantile

Φ : le niveau de confiance réellement atteint

$\lambda_0^{(j)} = 0,\ \forall j \in 1 \dots p$

for all $t \in 1 \dots T$ **do**

Estimation de $q^{(j)}_{\tau+\lambda_t^{(j)}}(\mathbf{x})$ selon (3), $\forall j \in 1 \dots p$

$\delta_i^{(j)} = Y_i^{(j)} - q^{(j)}_{\tau+\lambda_t^{(j)}}(\mathbf{x}_i),\ \forall i \in 1 \dots n, j \in 1 \dots p$

$\Phi = 1/n \sum_{i=1}^{n} \mathbb{1}_{\left[\sum_{j=1}^{p} \mathbb{1}_{[\delta_i^{(j)} > 0]} = p\right]}$

$S^{(j)} = 1/n \sum_{i=1}^{n} \mathbb{1}_{\left[\sum_{k \in \{1 \dots p\} \setminus j} \mathbb{1}_{[\delta_i^{(k)} > 0]} = p-1\right]},\ \forall j \in 1 \dots p$

$S^{(j)} = S^{(j)} / \sum_{k=1}^{p} S^{(k)},\ \forall j \in 1 \dots p$

$\lambda_{t+1}^{(j)} = \lambda_t^{(j)} + \eta(t) S^{(j)} (\Phi - 1 + \tau),\ \forall j \in 1 \dots p$

$\lambda_{t+1}^{(j)} = \min(\gamma, \lambda_{t+1}^{(j)}),\ \forall j \in 1 \dots p$

if $\lambda_{t+1}^{(j)} = \lambda_t^{(j)},\ \forall j \in 1 \dots p$ **then**

break

end if

end for

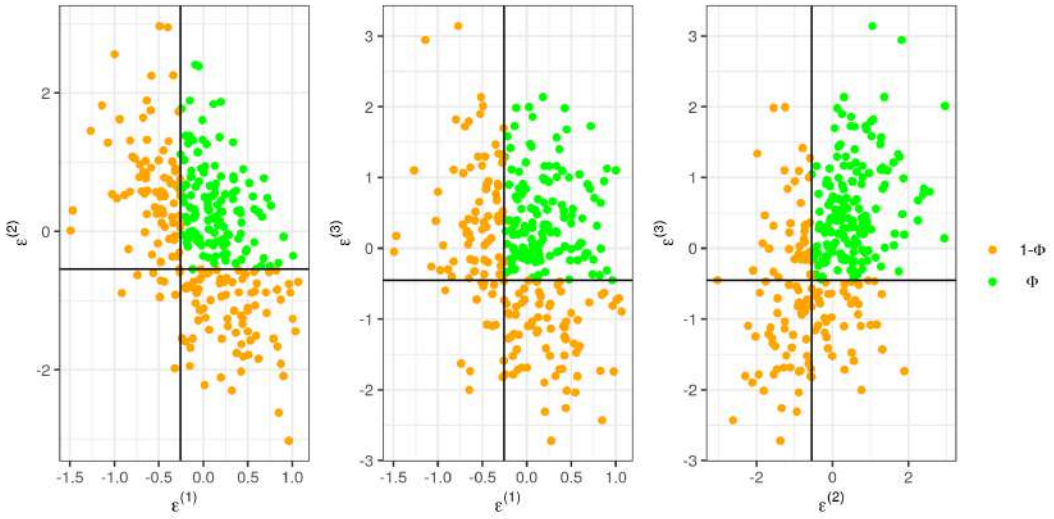


Figure 1: Niveau de confiance global Φ pour un risque $\tau = 0.3$ en fonction de la dépendance entre les erreurs aléatoires $\epsilon^{(j)}$: r=-0.48 entre $\epsilon^{(1)}$ et $\epsilon^{(2)}$, r=-0.21 entre $\epsilon^{(1)}$ et $\epsilon^{(3)}$, r=0.44 entre $\epsilon^{(2)}$ et $\epsilon^{(3)}$

4 Simulations

Afin d'illustrer notre approche et d'en vérifier le bon comportement numérique, $n = 300$ observations de $d = 4$ variables de décision et $p = 3$ objectifs ont été générées. Les variables explicatives stockées dans la matrice $\mathbf{X}$ de dimension $n \times d$ ont été générées selon une distribution uniforme $\mathcal{U}(-2, 2)$. Les 3 objectifs ont été construits via des combinaisons linéaires des variables explicatives auxquelles un bruit Gaussien a été ajouté :

$$y^{(j)} = \mathbf{X}\beta^{(j)} + \epsilon^{(j)}. \tag{10}$$

Les bruits $\epsilon^{(j)}$ ont été générés de manière dépendante (voir par exemple à la Figure 1). Les résultats présentés à la Table 1 montrent d'abord, de manière logique, que lorsque l'aversion au risque est forte (τ faible), la valeur optimale atteinte est plus modeste. Ensuite, en-dessous d'un certain niveau de risque (pour $\tau = 0.15$), $\tau + \lambda^{(j)}$ atteint le plancher, fixé à $\gamma = 0.1$ afin d'éviter d'estimer le quantile dans des zones peu fournies en observations, ceci empêchant alors d'atteindre le niveau de confiance $1 - \tau$ souhaité.

Perspectives

L'approche proposée permet de résoudre des problèmes d'optimisation "data-driven" avec une gestion multi-objectifs de l'incertitude des modèles de substitution via l'optimisation fondée sur des quantiles. Pour poursuivre ce travail, il serait intéressant d'étudier le

Table 1: Résumé de la résolution du problème d'optimisation pour trois risques τ (= 0.15, 0.30, 0.50) avec l'application ou pas de la méthode proposée d'ajustement via l'algorithme 1. Lorsqu'il n'y a pas d'ajustement, les $\lambda^{(j)}$ sont nuls. Les valeurs des trois objectifs ($y^{(1)}$, $y^{(2)}$, $y^{(3)}$) de chaque front de Pareto ont été moyennées ($\pm$ l'écart-type) de manière à rendre compte de la valeur optimale atteinte par les solutions pour les trois τ choisis.

τ	Méthode	Φ	$y^{(1)}$	$\tau + \lambda^{(1)}$	$y^{(2)}$	$\tau + \lambda^{(2)}$	$y^{(3)}$	$\tau + \lambda^{(3)}$
0.15	Non ajustée	0.62	5.5 $\pm$ 2.1	0.15	7.3 $\pm$ 2	0.15	1.8 $\pm$ 3.1	0.15
0.15	Ajustée	0.75	4.8 $\pm$ 2.2	0.1	7.4 $\pm$ 2	0.1	0.67 $\pm$ 3.1	0.1
0.30	Non ajustée	0.36	5.9 $\pm$ 1.9	0.3	7.8 $\pm$ 2.6	0.3	2.6 $\pm$ 3.1	0.3
0.30	Ajustée	0.70	4.6 $\pm$ 2.5	0.11	7.6 $\pm$ 1.9	0.11	0.72 $\pm$ 3.2	0.14
0.50	Non ajustée	0.14	6 $\pm$ 2.4	0.5	8.4 $\pm$ 2.5	0.5	3.1 $\pm$ 3.7	0.5
0.50	Ajustée	0.50	5.4 $\pm$ 1.9	0.16	7.5 $\pm$ 1.9	0.19	2.1 $\pm$ 3	0.25

comportement de la pénalité dans différents scénarios de corrélation des erreurs $\epsilon^{(j)}$, même si l'on infère que la pénalité $\lambda^{(j)}$ sera très probablement d'autant plus forte que les erreurs aléatoires $\epsilon^{(j)}$ sont négativement corrélées entre elles et que le nombre p d'objectifs est grand. Enfin, le calibrage des pénalités $\lambda^{(j)}$ est dépendante de l'échantillon d'apprentissage et il serait donc judicieux de trouver une stratégie pour optimiser ces pénalités de manière plus robuste, évitant ainsi une forme de sur-apprentissage.

Bibliographie

Bertsimas D. et Thiele A. (2006). *Models, methods, and applications for innovative decision making*. INFORMS.

Conanec A., Chavent M., Ellies-Oury M-P. et Saracco J. (2020). Une méthodologie computationelle pour faire de l'optimisation multi-objectifs en élevage de précision, *Actes des 52èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, pp. 212-219.

Deb K., Agrawal S., Pratap A. et Meyarivan T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *parallel problem solving from nature - PPSN VI, Berlin*, pp. 849-858.

Ehrgott M. (2005), *Multicriteria optimization*, Springer Science Business Media.

McBride K., Sundmacher K. (2019). Overview of surrogate modeling in chemical process engineering, *Chemie Ingenieur Technik*, 3, pp. 228-239.

A hidden Semi-Markov model for inferring the structure of migratory bird flyway networks

Marie-Josée Cros[1], Nathalie Peyrard[1], Régis Sabbadin[1], Ronan Trépos[1], Sam Nicol[2]

[1] *INRAE, UR MIAT, F-31320 Castanet-Tolosan, France*

[2] *CSIRO Brisbane, Australia*

Résumé. Il est difficile de mesurer la connectivité entre les différentes étapes des chemins migratoires empruntés par les oiseaux du fait de l'étendue des zones géographiques couvertes et du très grand nombre d'individus concernés. Plutôt que de se baser sur des suivis individuels, rarement disponibles, pour inférer la connectivité, une approche complémentaire consiste à utiliser des comptages observés en certains sites. Pour cela, nous modélisons les routes migratoires comme un réseau pondéré et nous présentons un modèle de semi-Markov caché des trajectoires (non observées) des oiseaux et des comptages observés. L'application exacte des algorithmes classiques d'inférence utilisant la vraisemblance n'est pas accessible pour ce modèle, étant donnée la taille de l'espace des états cachés. Nous proposons donc deux algorithmes pour l'estimation approchée : une version Monte-Carlo de l'algorithme EM et un algorithme ABC. Nous comparons la qualité des estimateurs obtenus sur des données simulées et des données réelles correspondant aux voies migratoires de l'est de l'Asie et de l'Australiasie pour une espèce d'oiseau.

Mots-clés. oiseaux migrateurs, comptages, Modèle de semi-Markov caché, ABC, MCEM, Far Eastern Curlew

Abstract. Measuring the connectivity of birds' stopovers on the migratory paths is challenging due to the wide geographic areas and vast numbers of individuals involved. A complementary approach to tracking individuals is to infer connectivity from count data at known aggregation sites. We model here a migratory flyway as a weighted network and we present a Hidden Semi Markov Model of the hidden birds trajectories and the observed counts. Exact application of estimation algorithms based on the likelihood for inferring the model parameters is not possible for even a small number of sites due to the high dimension of the hidden state. To overcome this, we derived two algorithms for approximated estimation: Monte Carlo Expectation-Maximisation and Approximate Bayesian Computation. We compare the quality of estimation of these two approaches on synthetic data and on a case study of a migratory shorebird in the East Asian-Australasian flyway.

Keywords. migration birds, counts, HSMM, ABC, MCEM, Far Eastern Curlew

1. Introduction

Every year, more than 50 million shorebirds migrate from overwintering habitat in Australasia to breeding grounds in Siberia. This migration is threatened by development pressures at stopover sites along the migratory route. Prioritising the conservation of critical bird habitat requires knowledge of the routes followed by birds but priorisation remains difficult without knowing how sites are connected.

We propose to use statistical modelling and efficient inference tools to determine the most likely network of migration routes based on observed counts at stopovers. In practice, available count data are noisy due to irregular collection intervals and detection errors. We model the network, noisy data and duration of stopover time at sites as a set of interacting Hidden Semi Markov Models (HSMM). In the model, individual birds sojourn in stopover sites for a period of time before moving to other sites according to an unknown multinomial distribution that we aim to estimate. For this kind of HSMM, exact application of existing estimation methods based on the likelihood is not possible for even a small number of sites due to the dimension of the hidden state. We designed two dedicated estimation algorithms for our model: Monte Carlo Expectation-Maximisation and Approximate Bayesian Computation. We present and compare the efficiency and quality of estimation of these approaches on synthetic data.

Then we illustrate their behavior on an applied case study using citizen science count data of the Far Eastern Curlew in the East Asian-Australasian Flyway. Far Eastern Curlews (*Numenius madagascariensis*) are the largest migratory shorebirds in the world, making an annual migration from their breeding grounds in Siberia, Mongolia and Kamchatka through east Asia to their predominantly Australian wintering grounds, before returning to breed (see Figure 1). The species is listed as critically endangered in Australia. Little is known about how Far Eastern Curlews use stopover sites [7] or how the individual sightings data can be extrapolated to the population level.

2. The flyway model

Let us consider the migration of a population of N birds over a set of I distinct stopovers (sites) over time. Sites are connected via some migration routes (oriented edges). We assume that birds do not fly backward, so we have an ordering of the sites such that if $i < j$ then a bird cannot fly from site j to site i. Therefore the set of all potential connections is given by the set of oriented edges from i to j for every $i < j$. The objective is to infer which edges are really used and with which relative importance.

We consider that each bird trajectory is modeled as a semi-Markov model [9] over a finite discrete time horizon $H = \{0, 1, 2, \ldots, T\}$, and that the N bird trajectories are independent. The state of a trajectory at a given time can be one of the I sites, or the state 'death' which corresponds to a bird who dies before time T. Then the elements of the semi-Markov model of a bird's trajectory are the followings:

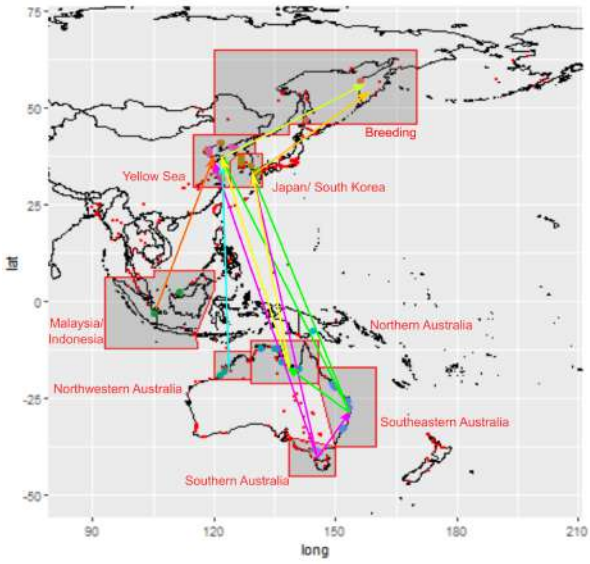


Figure 1: Map of case study network for the Far Eastern Curlew showing site boundaries and edges.

- Transition probabilities between states (the weights of the edges). The probability that any bird leaving site i at any given time goes to site j is $R(i,j)$. If $i \geq j$ then $R(i,j) = 0$. So R is an $I \times I$ upper triangular matrix. Note that, accounting for mortality, we may have, for any $i < I$, $\sum_{j \in 1 \ldots I} R(i,j) = \sum_{j=i+1}^{I} R(i,j) < 1$. The value $\mu_i = 1 - \sum_{j \in 1..I} R_{i,j}$, for $i < I$ is the mortality probability in site i which is assumed known. For a bird leaving site $i < I$, the destination is thus selected according to a multinomial distribution of parameters $(R(i,i+1), \ldots, R(i,I), \mu_i)$. For the breeding site I, we assume that when a bird 'leaves' site I, it necessarily moves to death so $\mu_I = 1$.

- Sojourn duration. We assume that the sojourn duration distribution in state $i \leq I$ is a shifted Poisson distribution of parameter λ_i. The shift is equal to one to ensure that sojourn duration is larger than 0. The sojourn duration in state 'death' is infinite ('death' is an absorbing state). Sojourn duration distributions depend on the site, but are the same for each bird. Furthermore, we assume that sojourn durations of two sites are independent.

In practice, the birds' trajectories are not observed and we infer the weights (the $R(i,j)$) of the network edges based on birds counts O_i^t for a set $\Omega \subseteq \{1 \ldots I\} \times \{1 \ldots T\}$ of observed site-times. These observed counts are noisy due to error detection. The real counts are the N_i^t, i.e. the number of birds located in site i at time t. We consider that, conditionally on the birds' trajectories, the observed counts are independent and the conditional distribution of O_i^t (for $(i,t) \in \Omega$) is modeld as a Poisson distribution, with parameter N_i^t. The model parameters are therefore $\theta = (R, \lambda_1, \ldots, \lambda_I)$.

3. Parameters estimation

Estimating the model parameters is not easy due to the following reasons. First, this is a model with hidden data: neither the individual bird trajectories nor the real counts N_i^t are observed. Second, conditionally on the observed counts, the N bird trajectories are no longer independent. We designed two simulation-based methods to estimate the parameters, based respectively on the Monte Carlo Expectation-Maximisation method (MCEM) [1] and the Approximate Bayesian Computation method (ABC) [3]. For MCEM, we used a Metropolis-Hasting algorithm to implement the E step. For ABC, we used the Lenormand sequential sampling method of the EasyABC package[1] in R [6]. We selected the set of all observed counts, O, as the summary statistics. Uniform priors were used for all parameters.

4. Results

4.1. Experiments on simulated data sets

We compared MCEM and ABC estimators on different simulated problems for different numbers of sites (from 4 to 10), network structures (3), and numbers of missing observations (0, 30 and 50 %). The 3 network structures correspond to 3 different values for the maximal number of destination sites per site (2, 3 and 4). For a given configuration (number of sites, structure, percentage of missing observations), 5 problems were generated. A problem is a vector of parameters θ and simulated trajectories (10 000 birds initially allocated randomly among source sites) and observed counts. The total number of problems is 300, since some configurations are not possible. For a given problem, MCEM was run with 5 different initialisations of the parameters and the best estimator in terms of likelihood was kept. The output of ABC is an estimate of the a posteriori distribution of the parameters. In order to compare ABC results with the true parameters values, we extracted the mean values of ABC marginals of each parameter. Finally we rescaled each parameter between 0 and 1 and we computed the mean absolute error (meanAE) between true and estimated rescaled parameters. Over the 315 problems, the mean value of meanAE was of 0.08 for ABC and 0.06 for MCEM. As expected we observed an increase of meanAE when the number of sites (see Figure 2) or the maximal number of neighbors increased, but in all cases the mean value of meanAE was always below 0.21. meanAE was not sensible to the percentage of missing observations, probably because bird's trajectories are smooth.

[1] https://CRAN.R-project.org/package=EasyABC, version 1.5

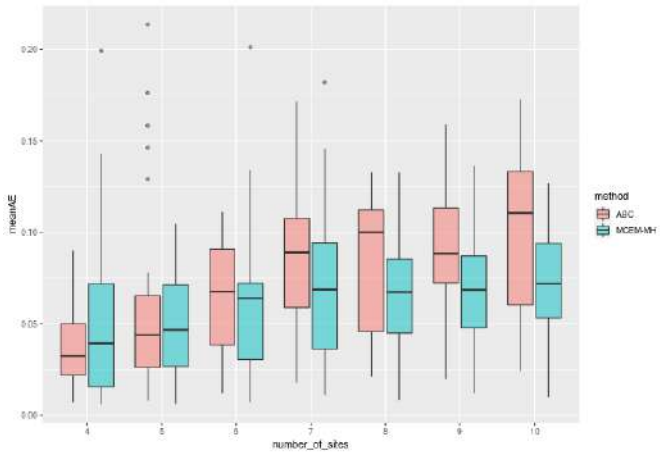


Figure 2: Overview of benchmarks results for methods ABC and MCEM, according to the number os sites.

4.2. Case study

We modeled the flyway using a network of 8 sites (Figure 1) representing the major known stopover regions for Far Eastern Curlews. Sites and edges are based on observed sightings and descriptions [7, 8], an expert-derived network [5] and eBird observations. The geographical extent of sites are defined to capture the important bird areas for Far Eastern Curlews [2]. Count data at each site is determined by combining the information of the eBird cheklists available in the site area, with post-treatment to reduce spatial bias. We applied the HSMM model on the eBird data from 2019. There are 32 000 birds with initial repartition derived from the observed count at time zero. The estimated networks are displayed on Figure 3. These are preliminary results, but as opposed to the experiments on simulated data, the estimators provided by ABC and MCEM were different. ABC results seem more coherent with expert knowledge. It may be more robust than MCEM for real data that deviate from the assumed HSMM model.

5. Conclusion

The HSMM model presented here is the first extension of Factorial HMM [4] to the hidden semi-Markov case and the first use for migratory network inference. The advantage of our approach is that it makes it possible to estimate flyway based only on limited count data at stopover sites. It enables the use of unstructured citizen science data. Preliminary results on the Far Eastern Curlew data are promising but additional analyses are needed. One limitation of the MCEM and ABC algorithms could be their computational time. To circumvent this problem, we are currently investigating a VBEM algorithm, which should be faster with, hopefully, a quality of estimation maintained.

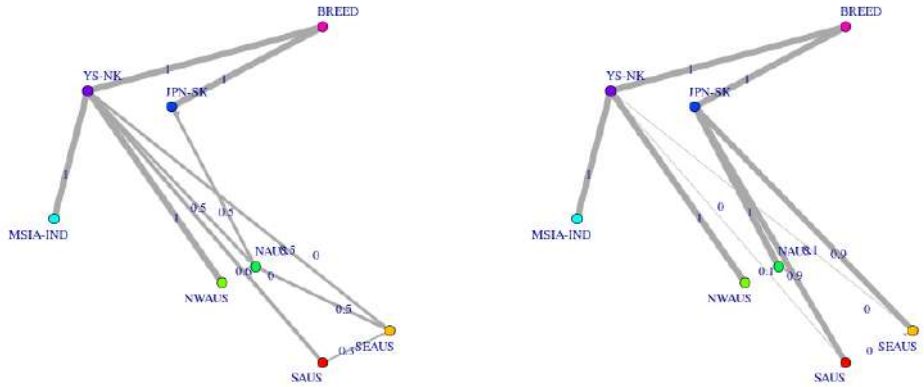


Figure 3: Estimated flyway network for the Far Eastern Curlew. Left: ABC; Right MCEM.

References

[1] C. Andrieu, N. De Freitas, A. Doucet, and M.L. Jordan. An introduction to MCMC for machine learning. Machine Learning, 50:5–43, 2003.

[2] M. Bamford, D. Watkins, W. Bancroft, G. Tischler, and J. Wahl. Migratory shorebirds of the east asia-australasia flyway: population estimates and internationally important sites. Technical report, Wetlands International– Oceania, 2008.

[3] M. G. B. Csilléry, K.and Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology and Evolution, 25(7):410–418, 2010.

[4] Z. Ghahramani and M. Jordan. Factorial hidden Markov models. Mach. Learn., 29(2-3):245–273, 1997.

[5] T. Iwamura, H. P. Possingham, I. Chadès, C. Minton, N. J. Murray, D. I. Rogers, E. A. Treml, and R. A. Fuller. Migratory connectivity magnifies the consequences of habitat loss from sea-level rise for shorebird populations. Proceedings of the Royal Society B: Biological Sciences, 280(1761):20130325, 2013.

[6] F. Jabot, T. Faure, and N. Dumoulin. Easyabc: performing efficient approximate bayesian computation sampling schemes using r. Methods in Ecology and Evolution, 4(7):684–687, 2013.

[7] C. Minton, R. Jessop, P C., and R. Standen. The migration of Eastern Curlew Numenius Madagascariensis to and from Australia. Stilt, 59:6–16, 2011.

[8] C. Minton, J. Wahl, H. Gibbs, R. Jessop, C. Hassell, and A. Boyle. Recoveries and flag sightings of waders which spend the non-breeding season in Australia. Stilt, 59:17–43, 2011.

[9] S.-Z. Yu. Hidden semi-Markov models. Artificial Intelligence, 2010.

Adapter la prise en charge des patients BPCO à leur profil: Classification de trajectoires physiologiques au cours du test de marche de 6 minutes en début de réadaptation chez les patients ayant une broncho-pneumopathie chronique obstructive

Mathieu DAVID, Jean-Yves DEGOS et Sébastien SAMYN

Patients with chronic obstructive pulmonary disease (COPD) can be classified into four groups based on heart rate or pulse oxygen saturation. Clustering of time series of heart frequency based upon Euclidean distance can be replaced by a clustering obtained from the mean and the relative standard deviation of the same times series (adjusted Rand index of 0; 95) without hardly any loss of information. It is slightly less true for the time series of the pulse oxygen saturation (adjusted Rand index of 0; 92 using the mean only). However, clusters based on statistics appear to be more strongly related to a significant improvement in the six-minutes walking test when using the 2 test.

The clusters are mainly determined by average levels of heart frequencies or pulse oxygen saturations. For each measure, we get an optimal partition in four groups. However, the partitions based on the two measures do not match : the patients with a low heart frequency are not necessarily those who keep a high pulse oxygen saturation. The best prediction of the significant progress is obtained by crossing the two partitions (namely 16 clusters). Risk seems important for patients with a high heart frequency, but those results should be confirmed with larger samples and are still uncertain, because they are obtained for small size groups.

SIMULATION ET IMPUTATION DE PLUSIEURS VARIABLES CORRELEES DANS UN CONTEXTE DE DONNEES MANQUANTES DE FAÇON NON ALEATOIRE (MNAR)

Joe DE KEIZER[1] & Antoine DUPUIS[2] & Sylvie RABOUAN[3] & Nicolas VENISSE[4] & Pascal CARATO[5] & Elise GAND[6] & Julie PAUL[7] & Emmanuelle BICHON[8] & Yoann DECEUNINCK[9] & Marion ALBOUY[10]

Direction de la recherche, Centre Hospitalier Universitaire de Poitiers, 2 rue de la Milétrie 86000 Poitiers, France, [1]joe.de-keizer@chu-poitiers.fr, [7]julie.paul@chu-poitiers.fr,

Groupe Health Endocrine Disruptors Exposome (HEDEX), INSERM-CIC1402, Centre Hospitalier Universitaire de la Milétrie, 2 rue de la Milétrie 86000 Poitiers, France, [2]antoine.dupuis@univ-poitiers.fr, [3]sylvie.rabouan@univ-poitiers.fr, [4]nicolas.venisse@chu-poitiers.fr, [5]pascal.carato@univ-poitiers.fr, [10]marion.albouy.llaty@univ-poitiers.fr

Centre d'investigation clinique CIC 1402, 2 rue de la Milétrie 86000 Poitiers France, [6]elise.gand@chu-poitiers.fr

LUNAM Université, Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), USC 1329, Oniris, 44307, Nantes, France, [8]emmanuelle.bichon@oniris-nantes.fr, [9]yoann.deceuninck@oniris-nantes.fr,

Résumé. Les dosages biologiques sont tous soumis à une limite de quantification (LOQ) des méthodes analytiques utilisées. Cela amène à analyser dans les études des variables contenant des données manquantes de façon non aléatoire (MNAR). Différents dosages issus des mêmes échantillons peuvent être à la fois corrélés entre eux et faire l'objet de données manquantes liées à une limite de quantification. L'objectif du projet est de réussir à simuler ce type de données manquantes afin d'étudier et de comparer les différentes techniques d'imputations simple et multiple proposées dans le cas de données manquantes MNAR.

Mots-clés. données manquantes, MNAR, corrélation, imputation

Abstract. Biological samplings are subject to a limit of quantification (LOQ) due to analytical methods used. This generates variables containing missing not at random data (MNAR). Different dosages from the same samples can be correlated with each other and subject to missing data. The objective of the project is to simulate these particular missing data to study and compare single and multiple imputation methods proposed for missing not at random data (MNAR).

Keywords. Missing data, MNAR, correlation, imputation

Les dosages biologiques sont tous soumis à une limite de quantification (LOQ) des méthodes analytiques utilisées. Cela amène à analyser dans les études des variables contenant des données manquantes de façon non aléatoire (MNAR). Dans un même échantillon plusieurs dosages peuvent être à la fois corrélés entre eux et faire l'objet de données manquantes liées à une limite de quantification.

Dans le cas où ces données manquantes sont nombreuses dans les sets de données, l'analyse peut difficilement se faire sur les « cas complets ». Il est alors envisagé d'utiliser des méthodes d'imputation. Or, il n'y a pas, à notre connaissance, de méthodes d'imputation de données manquantes MNAR spécifiquement adaptées à des sets de données contenant des variables à imputer corrélées entre elles. Nous proposons dans ce travail de simuler et de tester différentes techniques d'imputation simple et multiple sur plusieurs variables corrélées présentant des données manquantes MNAR.

Les données, qui ont amené la réflexion de notre groupe, sont celles issues de la cohorte EDDS.

1. La cohorte EDDS

Le protocole de recherche clinique «Exposition hydrique aux perturbateurs endocriniens des femmes enceintes en Deux-Sèvres : comparaison de méthodes d'évaluation et relation avec la santé de l'enfant » a donné lieu à la création d'une base de données, la cohorte EDDS.
Ce protocole de recherche biomédicale a reçu un avis favorable du Comité de protection des personnes Ouest III. Chaque participante a donné son consentement pour le recueil de données médicales les concernant. Entre 2012 et 2013, 144 femmes enceintes vivant dans les Deux-Sèvres ont été incluses.

1.1. Les objectifs

L'objectif de la recherche EDDS est de comparer trois méthodes validées et non invasives d'estimation hydrique de l'exposition aux perturbateurs endocriniens afin d'identifier celle qui est la plus précise pour estimer les apports réels. Il s'agit d'utiliser la méthode des triades à partir du recueil de :

- La mesure de biomarqueur d'exposition (dosage dans le colostrum),
- D'un questionnaire couplé au monitoring de l'environnement (dosage en sortie d'usine de traitement de l'eau)
- Du même questionnaire couplé à un monitoring individuel (dosage de l'eau du robinet du domicile).

Pour réaliser cette évaluation individuelle, chaque femme incluse dans le protocole a bénéficié de deux visites à domicile, une au cours du deuxième trimestre de grossesse et une au cours du troisième trimestre. A chacune de ces visites des échantillons d'eau étaient collectés au robinet du domicile des personnes. Des questionnaires d'exposition étaient également proposés aux femmes incluses lors de ces entretiens. Grâce aux données collectées Albouy-Llaty (2015) a pu estimer l'exposition à l'eau des femmes enceintes en prenant en compte l'eau ingéré et l'exposition cutanée.
Afin d'évaluer l'axe de monitoring individuel de la triade, il était nécessaire de mettre cette mesure individuelle de l'exposition à l'eau en lien avec le taux de perturbateurs endocriniens présent dans l'eau de consommation du domicile des femmes.

Le perturbateur endocrinien spécifiquement étudié par le protocole clinique est le bisphénol A (BPA), produit en grande quantité par l'industrie du plastique. Il s'agit d'un perturbateur retrouvé dans les eaux de consommation. Malgré le traitement des eaux de surface par les usines de traitement d'eau potable, le BPA est retrouvé en sortie d'usine. De plus, le traitement réalisé par ces usines, et plus particulièrement la chloration, génère des dérivés chlorés spécifiques du BPA qui possèdent une activité perturbatrice endocrinienne 100 fois plus importante que les molécules mères.

Le risque d'exposition humaine est élevé, d'autant que les perturbateurs endocriniens pourraient induire des effets même à faible dose. Ce risque peut être transgénérationnel car la grossesse est une période critique durant laquelle le fœtus a une plus forte susceptibilité aux œstrogènes. Des études animales ont montré une relation entre exposition aux perturbateurs endocriniens pendant la gestation et surpoids de la portée. La cohorte EDDS cherche donc à obtenir une estimation précise de l'exposition hydrique aux BPA et ses dérivés chlorés.
Le dosage des perturbateurs endocriniens a été réalisé par le LABERCA avec un objectif, à ce jour, d'amélioration de la limite de quantification.

1.2. Les données disponibles dans la cohorte

Les échantillons d'eau à destination de consommation humaine recueillis à domicile au cours du second trimestre de grossesse des femmes de la cohorte EDDS sont disponibles pour 99 d'entre elles. Chacun des échantillons d'eau est dosé afin de définir sa teneur en BPA et en dérivés chlorés sous les formes mono, bi1, bi2, tri ou tétrachlorées. Les teneurs en dérivés chlorés sont comprises respectivement entre [0,09 - 59,21 ng.L-1], [0,17 - 127,20 ng.L-1], [0,16 - 42,19 ng.L-1], [0,13 - 4,32 ng.L-1] et [0,09 - 2,28 ng.L-1] pour les formes mono, bi1, bi2, tri et tétrachlorées.
Le dosage de ces dérivés donne lieu à la création de données manquantes. En effet, pour respectivement 38%, 68%, 70%, 87% et 85% des échantillons, les valeurs des dérivés chlorés du BPA (mono, bi1, bi2, tri et tétrachlorées) sont en dessous d'une limite de quantification.
En fonction des kits utilisés, les LOQ ne sont pas les mêmes pour l'ensemble des échantillons.
Les dérivés chlorés de BPA étant présents conjointement dans les échantillons, nous étudions la corrélation excitante entre les taux estimés dans les échantillons d'eau. Une forte corrélation entre les dérivés chlorés est retrouvée (jusqu'à un rho de Spearman de 0.96 entre les formes bi2 et tri). Nous souhaitons recréer des sets de données présentant des proportions de données manquantes MNAR et une corrélation proche de ces constatations.

1.3. La simulation des variables corrélées avec des données manquantes MNAR

La simulation porte sur la création de 100 sets de 100 dosages chacun, contenant cinq variables continues avec un taux de données manquantes respectivement de 30%, 45%, 60%, 75% et 90%. La corrélation de Spearman simulée entre les différentes variables est comprise entre [0,30 ; 0,85]. Chacune des cinq variables simulées a une moyenne et un écart type spécifiques fixés à l'avance, associés à l'exponentielle d'une loi normale.

Table 1 : variables simulées et les paramètres associées

Variable	Paramètres loi normale	Pourcentage DM en % (±3%)
V1	μ=0,1 σ=1,8	30
V2	μ=-0,6 σ=1,6	45
V3	μ=-1,2 σ=1,4	60
V4	μ=-1,7 σ=1,2	75
V5	μ=-2,2 σ=1,0	90

Ces paramètres sont choisis afin de se rapprocher le plus possible des données de la cohorte EDDS et afin d'avoir les taux de données manquantes recherchées pour chacune des cinq variables.
Pour chaque variable, les valeurs simulées sont associées aléatoirement à une LOQ parmi les quatre seuils suivant : 0,2, 0,4, 0,6 ou 0,8. Les données des sets inférieurs à la LOQ associée sont considérées comme des données manquantes.
La médiane (Q1 – Q3) des itérations nécessaires à la création de ces sets de données est de 2028 (902 – 3438) pour un temps médian machine de 6,5 (2,8 – 10,8) secondes.
L'ensemble des analyses est réalisé à l'aide du logiciel R version 4.0.2.

2. Les méthodes d'imputation

2.1. Les différentes techniques proposées dans le cas de données manquantes MNAR

Les méthodes d'imputation sélectionnées pour leur possible utilisation dans le contexte de données manquantes MNAR sont :

- L'imputation simple par la moitié de la valeur de LOQ (HM),
- Quantile Regression Imputation of Left-Censored data (QRILC),
- Multiple Imputation by Chained Equations (MICE),
- Below LOQ using censored maximum likehood (BLOQ),
- Gibbs Sampler based left-censored missing value imputation approach (GSimp),
- Truncation k-Nearest Neighbor imputation (kNN-TN).

HM, half minimum (Little & Rubin, 1987), est une méthode d'imputation simple qui consiste à remplacer les données manquantes par la moitié de la plus petite valeur observée.
QRILC (Lazar, 2016), est une méthode qui impute les données manquantes dues à une valeur inférieure à la limite de détection par tirage au hasard des valeurs d'une distribution tronquée estimée par une régression quantile.
MICE, ou imputation multiple par équations enchainées (Azur, 2011), est une méthode d'imputation pour données manquantes MAR (missing at random) qui sélectionne un modèle d'imputation différent pour chaque type de variable. C'est une méthode itérative, c'est-à-dire qu'elle impute les données manquantes une nouvelle fois à chaque itération en se basant sur les résultats obtenus à l'itération précédente. Pour nos variables continues nous avons utilisé un modèle PMM (predictive mean matching).
La BLOQ regroupe différentes méthodes d'imputation, celle dont nous nous intéressons est l'imputation par le maximum de vraisemblance censuré (Barnette, 2020). Cette méthode combine l'estimation supérieure de la moyenne et de la variance qu'apporte le maximum de vraisemblance censuré et conserve la structure de la relation entre les points utilisée par les méthodes d'imputation. Toutes les valeurs imputées sont en dessous de la plus grande LOQ.
GSimp, ou Gibbs sampler (Wei, 2018), est une méthode d'imputation par une technique de Monte Carlo à chaîne de Markov qui permet de mettre à jour les observations alors que certaines sont fixes. Toutes les valeurs imputées sont comprises entre 0 et la plus petite valeur observée.
KNN-TN (Shah, 2017), est une méthode d'imputation basée sur la recherche du plus proche voisin, qui tient compte de la troncature à la LOQ la plus élevée. La troncature intervient lorsqu'il n'est pas possible de connaitre les données qui se situent en dessous du seuil qui a été défini.

Ces méthodes sont comparées à l'aide de l'indicateur de performance d'erreur quadratique moyenne normalisée (NRMSE). Le NRMSE est utilisé pour évaluer la précision en calculant les différences entre les valeurs imputées et les valeurs réelles, pour ainsi déterminer la méthode la plus pertinente, selon la formule :

$$NRMSE = \frac{\sqrt{\mu((X^T - X^I)^2)}}{\sigma(X^T)} \quad (1)$$

Où X^T correspond aux valeurs réelles et X^I aux valeurs imputées.
Plus les valeurs de l'indicateur sont basses et plus les méthodes semblent performantes pour imputer des données manquantes.

Pour des données manquantes de type MNAR, l'utilisation du NRMSE pourrait entraîner des résultats biaisés comme le montre l'article de Wei (2018). Par conséquent une autre mesure a été

dérivée, la somme des rangs (SOR) basée sur le NRMSE. C'est une mesure non paramétrique, qui permet de comparer les différentes méthodes d'imputation de façon robuste et non biaisée.

$$SOR = \sum_{i=1}^{k} Rank_i(NRMSE) \quad (2)$$

Où k est le nombre de variables manquantes et $Rank_i$(NRMSE) le rang de NRMSE des différentes méthodes d'imputation dans la $i^{ème}$ variable manquante.

2.2. Imputation sur les données simulées et comparaison

Pour chacun des 100 sets de données simulées, les six méthodes d'imputation sont répétées 50 fois afin d'augmenter la précision. Les résultats de ces imputations itératives sont comparés par les indicateurs NRMSE et SOR.

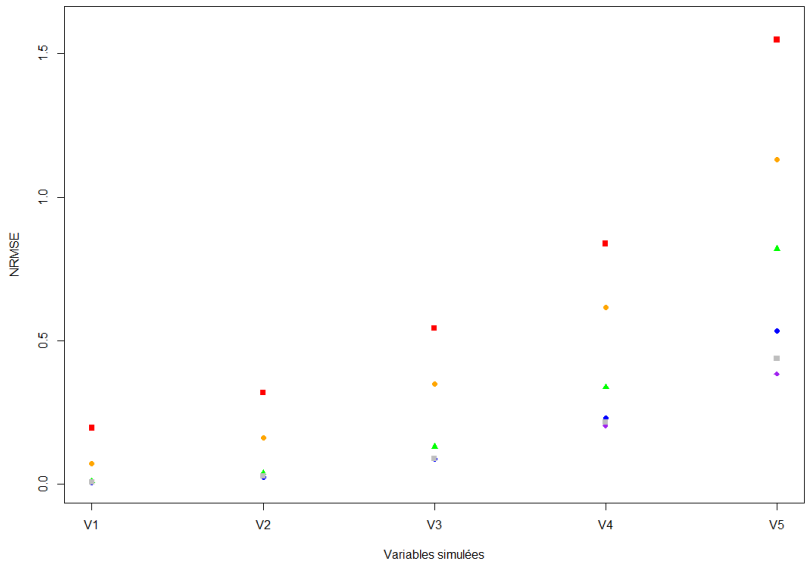


Figure 1. Moyenne des NRMSE pour les cent sets de données simulées pour chacune des variables simulées (V1-V5) en fonction des techniques d'imputation : HM (gris), QRILC (bleu), MICE (rouge), BLOQ (vert), GSimp (violet) et trKNN (orange).

Les méthodes HM, QRILC et GSimp montrent la plus faible différence entre les données simulées et celles imputées pour les différents variables avec des taux de données manquantes différents.
Leurs performances évaluées par NRMSE pour ces trois méthodes d'imputation sont proches.

Nous retrouvons des résultats identiques en considérant l'indicateur SOR. Les moyennes ± écart type des résultats pour les 100 sets de données simulées sont les suivants : HM 9,85 ± 0,94, QRILC 9,97 ± 1,19, MICE 29,81 ± 0,44, BLOQ 20,07 ± 0,73, Gsimp 10,35 ± 1,20 et trKNN 24,95 ± 0,63.

3. Conclusion

Les techniques itératives permettent de simuler des variables corrélées les unes aux autres en présence de données manquantes MNAR.
Les méthodes HM, QRILC et GSimp semblent les plus performantes pour imputer des variables contenant ce type de données manquantes surtout lorsque le taux de données manquantes est important.
Les perspectives de travail sont de tester la robustesse de ces résultats pour des jeux de données de taille plus importante, de prendre en compte la présence de plusieurs LOQ au sein d'une même variable, de considérer plusieurs variables d'ajustement telles que la distance entre le domicile des

femmes enceintes et l'usine de traitement des eaux ou encore la présence à domicile d'appareil de traitement de l'eau (adoucisseur par exemple), dans les techniques d'imputation qui le permettent.

Bibliographie

Albouy-Llaty M, Dupuis A, Grignon C, Strezlec S, Pierre F, Rabouan S, Migeot V. (2014) Estimating drinking-water ingestion and dermal contact with water in a French population of pregnant women: the EDDS cohort study *Journal of Exposure Analysis and Environmental Epidemiology*

Little R.J.A., Rubin D.B. (1987) Statistical analysis with missing data, *Wiley series in probalility and statistics*

Lazar C, Gatto L, Ferro M, Bruley C, Burger T. (2016) Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research*

Wei R, Wang J, Jia E, Chen T, Ni Y, Jia W. (2018) GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies *PLOS Computational Biology*

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. (2011) Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research.*

Barnett HY, Geys H, Jacobs T, Jaki T (2020) Methods for Non-Compartmental Pharmacokinetic Analysis With Observations Below the Limit of Quantification, *Statistics in Biopharmaceutical Research*

Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. (2017) Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics*. 18(1):114

Modèles à effets mixtes pour l'inférence de dynamiques épidémiques multisites.

Romain Narci [1] & Maud Delattre [2] & Catherine Larédo [3] & Elisabeta Vergu [4]

MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[1] *romain.narci@inrae.fr ;* [2] *maud.delattre@inrae.fr ;* [3] *catherine.laredo@inrae.fr ;*
[4] *elisabeta.vergu@inrae.fr*

Résumé. L'estimation des paramètres régissant les phénomènes épidémiques à partir des données disponibles est un enjeu majeur, notamment pour mieux comprendre les mécanismes à la base de ces dynamiques et en fournir des prédictions fiables. Ces données sont systématiquement entachées de différentes sources de bruit (erreurs de mesure, bruits d'observation, sous-déclaration, etc.), avec de plus certaines coordonnées décrivant ces dynamiques qui ne sont pas observées. Pour améliorer l'analyse statistique, nous proposons d'étudier dans un modèle unique les dynamiques épidémiques se produisant simultanément dans différentes régions. Chaque épidémie est modélisée par un processus stochastique caractérisant une variabilité intrinsèque et la variabilité inter-régions est prise en compte par l'inclusion d'effets aléatoires sur les paramètres régissant ces dynamiques. L'estimation par maximum de vraisemblance requiert l'utilisation d'algorithmes stochastiques. Chaque dynamique est modélisée par un processus Gaussien à petite variance. Cette approximation permet de coupler l'algorithme SAEM à des techniques basées sur le filtre de Kalman afin d'estimer les paramètres épidémiques. Les performances de l'algorithme ainsi construit sont étudiées à l'aide de simulations de dynamiques SIR.

Mots-clés. Modèles à effets mixtes, algorithme SAEM, filtre de Kalman, dynamiques épidémiques, inférence des paramètres du modèle.

Abstract. The estimation of parameters governing epidemic phenomena from available data is a major challenge, especially to better understand the mechanisms underlying these dynamics and to provide reliable predictions. These data are systematically affected by various sources of noise (measurement errors, observation noises, under-reporting, etc.), with in addition certain coordinates describing these dynamics that are not observed. To improve the statistical analysis, we propose to study in a unique model the epidemic dynamics occurring simultaneously in different regions. Each epidemic is modeled by a stochastic process characterizing an intrinsic variability and inter-regional variability is taken into account by including random effects on the parameters governing these dynamics. The maximum likelihood estimation requires the use of stochastic algorithms. Each dynamic is modeled by a Gaussian process with small variance coefficient. This approximation allows to couple the SAEM algorithm with Kalman filtering based techniques in order to estimate the epidemic parameters. The performances of the algorithm are studied using simulations of SIR dynamics.

Keywords. Mixed effects models, SAEM algorithm, Kalman filter, epidemic dynamics, inference of model parameters.

1 Introduction

Contexte Proposer une modélisation mathématique pertinente et des méthodes statistiques rigoureuses est un enjeu majeur en épidémiologie des maladies infectieuses. Les données disponibles dans ce cadre sont systématiquement entachées de différentes sources de bruit (erreurs de mesure, bruits d'observation, sous-déclaration, etc.), et certaines coordonnées décrivant ces dynamiques ne sont pas observées. De plus, les dynamiques épidémiques peuvent être récurrentes dans le temps et/ou se produire simultanément dans différentes régions. A titre d'exemple, la grippe humaine en France est saisonnière et les épidémies peuvent se déployer dans plusieurs régions distinctes avec des intensités différentes au même moment. En pratique, cette variabilité est souvent gommée en omettant de prendre en compte de manière explicite la composante spécifique à chaque entité (population, période) ou est estimée de façon plus ou moins empirique ou alors de manière séparée pour chaque série de données. Intégrer directement dans un modèle ces sources de variabilité permettrait d'améliorer la puissance statistique et la précision de l'estimation des paramètres épidémiques ainsi que de leur variabilité, en considérant simultanément les jeux de données observés correspondant à chaque entité spatiale (e.g. région) ou temporelle (e.g. saison).

Les modèles à effets mixtes permettent de décrire une variabilité entre sujets d'une même population à partir de données répétées. Ils sont notamment utilisés en pharmacocinétique avec des dynamiques intra-population modélisées par des équations différentielles ordinaires (EDO), et des effets aléatoires sur les paramètres de ces dynamiques pour décrire les différences entre individus. A notre connaissance, ce cadre des modèles à effets mixtes a été peu utilisé pour les dynamiques épidémiques. [Bretó et al., 2020] se sont intéressé à des données de panel issues de modèles non linéaires partiellement observés dans le cadre d'études épidémiologiques et ont proposé une méthode d'inférence basée sur la vraisemblance utilisant des techniques de filtrage particulaire. [Prague et al., 2020] ont analysé les données de la première vague épidémique de COVID-19 en France à l'aide d'un système d'équations différentielles ordinaires incorporant des paramètres aléatoires pour tenir compte de la variabilité des dynamiques entre régions.

Les objectifs de notre travail sont

1. proposer une modélisation plus fine des épidémies multisites grâce aux modèles à effets mixtes, en y incluant une modélisation stocahstique de chaque épidémie plutôt qu'une modélisation déterministe de type EDO,
2. développer une méthode appropriée pour en estimer les paramètres.

Pour décrire les dynamiques épidémiques, nous utilisons la formulation autorégressive obtenue à partir de processus Gaussiens à petite variance que nous avons étudiée dans un précédent travail [Narci et al., 2020]. Ce modèle, facilitant l'inférence de paramètres, est utilisé et étendu au cadre des modèles à effets mixtes. Nous proposons alors de coupler l'algorithme SAEM avec des techniques basées sur le filtre de Kalman afin d'estimer les paramètres épidémiques dans ce modèle. Les performances de l'algorithme ainsi construit sont étudiées sur données simulées.

2 Modèle

Nous partons d'une modélisation compartimentale permettant de décrire l'évolution au cours du temps du nombre d'individus dans différents stades de santé vis-à-vis du pathogène. Comme exemple d'application pour l'écriture du modèle et pour les études numériques, nous utilisons le modèle épidémique simple SIR décrit dans la Figure 1, l'extension à des modèles plus complexes (SEIR, SEIRS, etc.) s'en déduisant facilement.

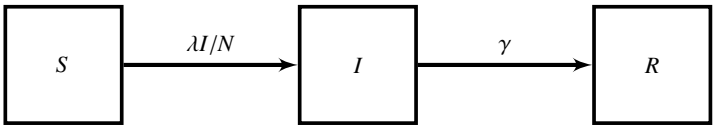


FIGURE 1 – Modèle à compartiments SIR avec trois blocs correspondant respectivement à des individus susceptibles (S), infectés (I) et guéris (R). Les transitions des individus d'un état de santé à un autre sont régies par le taux de transmission λ et le taux de guérison γ.

Considérons maintenant la situation où une même épidémie se produit dans plusieurs régions simultanément. Nous utilisons l'indice $1 \leq u \leq U$ pour décrire les quantités relatives à chaque région, où U est le nombre total de régions. Ainsi, la dynamique épidémique de la région u est représentée par le processus Gaussien d-dimensionnel $(X_u(t))_{t\geq 0}$, $1 \leq u \leq U$, décrivant une épidémie à $d+1$ stades d'infection (ou compartiments) et d'espace d'états $E = \{0, \ldots, N_u\}^d$ où N_u est la taille de population dans la u-ème région. Chacune de ces dynamiques possède des paramètres qui lui sont propres, notés ϕ_u.

Dans le cas du modèle SIR, $d = 2$ et $X_u(t) = (S_u(t), I_u(t))^t$. Les paramètres spécifiques à une épidémie sont les taux de transition λ_u et γ_u et les proportions initiales de susceptibles et d'infectés $s_{0,u} := S_u(0)/N_u$ et $i_{0,u} := I_u(0)/N_u$. Dans les études de simulation, nous utilisons plutôt la reparamétrisation permettant d'exhiber les paramètres épidémiques d'intérêt $R_{0,u} = \lambda_u/\gamma_u$ (nombre de reproduction de base) et $d_u = 1/\gamma_u$ (durée d'infectiosité d'un individu infecté).

Soit $(X_u(t_k), t_k = k\Delta, k = 0, \ldots, n_u)$ le u-ème processus Gaussien épidémique discrétisé en temps, où n_u est le nombre d'observations pour l'épidémie u et Δ est le pas de temps, se déroulant sur un intervalle de temps fixe $[0, T_u]$, avec $T_u = n_u\Delta$. Dans [Narci et al., 2020], nous avons aussi proposé une écriture Gaussienne des observations partiellement observées et bruitées, que nous notons $Y_{u,k}$ pour une épidémie u et un temps d'observation t_k. Pour décrire le modèle à effets mixtes, nous utilisons une représentation à deux niveaux correspondant à deux types de variabilité.

1. Variabilité intra-épidémie Pour tout u, notons $X_{u,k} := X_u(t_k)/N_u$ et $X_{u,0} = x_{u,0}$. Conditionnellement à $\phi_u = \varphi$, nous avons le modèle suivant :

$$\begin{cases} X_{u,k} = F_k(\varphi) + A_{k-1}(\varphi)X_{u,k-1} + V_{u,k}, \ k \geq 1, \ 1 \leq u \leq U, \\ Y_{u,k} = B(\varphi)X_{u,k} + W_{u,k}, \end{cases} \tag{1}$$

où $V_{u,k} \sim \mathcal{N}_d(0, T_k(\varphi, \Delta))$ et $W_{u,k} \sim \mathcal{N}_q(0, Q_k(\varphi))$ avec F_k, A_{k-1}, T_k, B et Q_k connus. Les paramètres régissant les dynamiques épidémiques à estimer sont contenus dans A_k et T_k. L'opérateur B est un opérateur de projection traduisant le fait que des coordonnées ne sont pas observées (observations partielles). Nous supposons de plus que l'équation des observations $Y_{u,k}$ incorpore un bruit de mesure dépendant de u (modélisant un taux de report du nombre de personnes infectées différent selon les régions par exemple).

Dans le cas du modèle SIR, nous observons une proportion bruitée du nombre d'infectés (correspondant à une seule coordonnée du système, ici $I(t_k)$) à des temps discrets. L'opérateur B est la matrice de projection $B(\varphi) = (0 \quad p)$, où p est le taux de report du nombre de personnes infectées, et la matrice Q_k dépend aussi de p. Ainsi que les autres paramètres, le taux de report peut être aléatoire et noté p_u.

2. Variabilité inter-épidémie Pour modéliser la variabilité inter-épidémie, nous supposons que les paramètres du modèle ϕ_u sont aléatoires et vérifient :

$$\begin{cases} \phi_u &= h(\beta, c_u, \xi_u), \\ \xi_u &\sim \mathcal{N}(0, \Omega), \end{cases} \tag{2}$$

où h est connue, β est le vecteur des effets fixes, c_u sont d'éventuelles covariables définies pour une épidémie u et Ω est une matrice de covariance. Les paramètres inconnus du modèle, à estimer, sont $\theta = (\beta, \Omega)$.

Dans le cas du modèle SIR, les effets aléatoires sont $\phi_u = (R_{0,u}, d_u, p_u, s_{0,u}, i_{0,u})$.

3 Estimation par maximum de vraisemblance

Dans le modèle décrit par (1) et (2), l'estimation des paramètres est difficile car tous les stades du processus d'infection et tous les compartiments épidémiques ne sont pas observés. De plus, les effets aléatoires ϕ_u spécifiques à chaque épidémie sont inconnus. Ainsi, le calcul de l'estimateur du maximum de vraisemblance n'est pas explicite. Nous proposons d'adapter l'algorithme SAEM-MCMC [Kuhn and Lavielle, 2004] en utilisant des outils analogues à ceux de [Delattre and Lavielle, 2013]. Le point clé est la simulation des variables latentes du modèle à chaque itération de l'algorithme. En utilisant un filtre de Kalman exact dans le modèle (1), il est possible de marginaliser par rapport aux états épidémiques non observés X_u et de ne simuler que les ϕ_u. Les différences principales avec l'approche de [Delattre and Lavielle, 2013] proviennent du modèle utilisé pour décrire la variabilité intrinsèque à chaque dynamique, incluant ici un terme de petite variance, et de l'utilisation d'un filtre de Kalman exact plutôt qu'approché.

4 Résultats numériques

Les performances de cette méthode d'inférence ont été étudiées à partir de simulations de dynamiques SIR. Pour la suite, nous supposons que $R_{0,u}$ et p_u sont aléatoires, que $d_u = d$ est fixe et que

$s_{0,u} + i_{0,u} = 1$ (ce qui revient à considérer qu'aucun individu n'est immunisé lorsque débute l'épidémie) avec $i_{0,u}$ aléatoire. Les paramètres de chaque dynamique épidémique $\phi_u = (R_{0,u}, d_u, p_u, i_{0,u})^\top$, $u = 1, \ldots, U$, sont définis par

$$\phi_u = \beta + \xi_u,\ \xi_u \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega),$$

avec $\beta = \left(R_{0,pop}, d_{pop}, p_{pop}, i_{0,pop}\right)^\top$ et $\Omega = \text{diag}\left(\omega^2_{R_0}, 0, \omega^2_p, \omega^2_{i_0}\right)$. Les valeurs choisies dans l'expérience de simulation pour les paramètres du modèle sont : $\beta = (3, 3, 0.78, 0.11)^\top$ et
$\Omega = \text{diag}\left[(0.17 * R_{0,pop})^2, 0, (0.18 * p_{pop})^2, (0.45 * i_{0,pop})^2\right]$.

Cent jeux de données ont été simulés avec U épidémies, $U = \{20, 100\}$, évoluant chacune dans une population de taille $N_u = 10000$, $u = 1, \ldots, U$. Les observations, correspondant à une proportion bruitée du nombre des infectés, sont générées à des temps régulièrement espacés $t_k = k\Delta$ en utilisant la même valeur de Δ pour chacune des U épidémies. Comme la durée épidémique est stochastique, des intervalles d'observation spécifiques $[0, T_u]$ ont été construits pour chaque épidémie u et la valeur de T_u a été choisie comme le premier instant où le nombre d'individus infectés devient nul. Nous avons choisi $\Delta = 0.75$ de façon à considérer un nombre d'observations proche de 50 pour chaque dynamique épidémique. La figure suivante illustre la variabilité inter-épidémies pour les processus épidémiques Gaussiens ainsi que pour les observations :

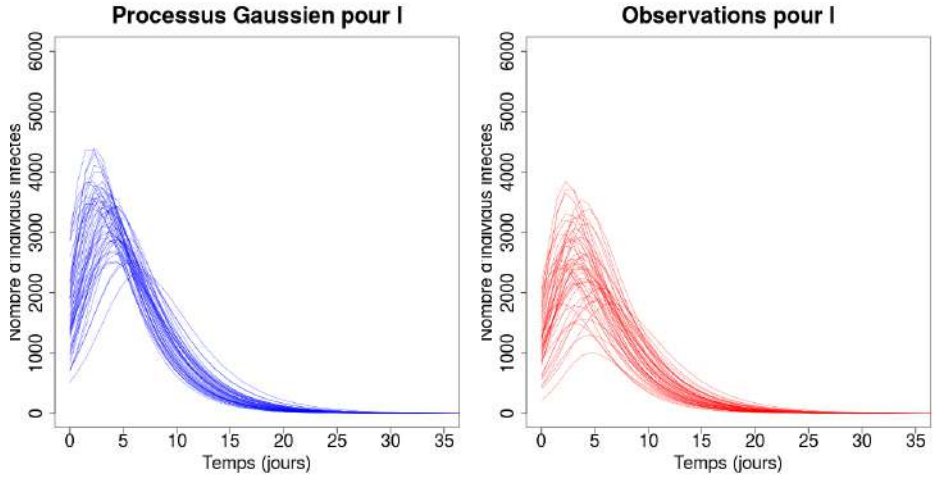


FIGURE 2 – Représentation de $U = 50$ trajectoires épidémiques. Graphique de gauche : 50 processus Gaussiens des I. Graphique de droite : 50 trajectoires épidémiques observées selon un modèle Gaussien des observations.

Pour chaque valeur de U, 100 estimations ponctuelles ont été obtenues par notre méthode d'inférence. Les résultats sont présentés dans le tableau (1).

TABLE 1 – Estimation de $\theta = (\beta, \Omega) = (R_{0,pop}, d_{pop}, p_{pop}, i_{0,pop}, \omega^2_{R_0}, \omega^2_p, \omega^2_{i_0})$ avec les vraies valeurs des paramètres θ^* indiquées sur la deuxième ligne. Pour chaque U et pour chaque paramètre du modèle, les estimations moyennes et les écarts-type sont calculés sur 100 estimations ponctuelles.

Paramètres (θ)		$R_{0,pop}$	d_{pop}	p_{pop}	$i_{0,pop}$	ω_{R_0}	ω_p	ω_{i_0}
Vraies valeurs (θ^*)		**3**	**3**	**0.78**	**0.11**	**0.5**	**0.14**	**0.05**
$U = 20$	Mean	3.037	3.018	0.780	0.108	0.605	0.138	0.046
	SD	0.227	0.149	0.048	0.012	0.228	0.030	0.010
$U = 100$	Mean	3.088	3.044	0.772	0.109	0.688	0.142	0.048
	SD	0.118	0.072	0.022	0.005	0.125	0.014	0.004

Les résultats montrent que les estimations sont quasiment sans biais même pour un nombre d'épidémies modéré ($U = 20$). Passer de $U = 20$ à $U = 100$ permet d'améliorer la précision des estimations en faisant diminuer de presque de moitié les écarts-type des estimations.

5 Conclusion générale

La méthode d'inférence que nous proposons dans le cadre des modèles à effets mixtes de dynamiques épidémiques peut être étendue à d'autres modèles à compartiments et prendre en compte des covariables dans la procédure d'estimation. La représentation à deux niveaux décrivant les variabilités inter-épidémie et intra-épidémie est très flexible dans la mesure où une liberté de choix est donnée pour la structure des observations, la nature des bruits d'observation et de mesure et l'incorporation ou non d'effets fixes/aléatoires dans les différents paramètres du modèle. Une application sur données réelles est actuellemement en cours.

Références

[Bretó et al., 2020] Bretó, C., Ionides, E., and King, A. (2020). Panel data analysis via mechanistic models. Journal of the American Statistical Association, 115(531) :1178–1188.

[Delattre and Lavielle, 2013] Delattre, M. and Lavielle, M. (2013). Coupling the saem algorithm and the extended kalman filter for maximum likelihood estimation in mixed-effects diffusion models. Statistics and Its Interface, 6 :519–532.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM : Probability and Statistics, 8 :115–131.

[Narci et al., 2020] Narci, R., Delattre, M., Larédo, C., and Vergu, E. (2020). Inference for partially observed epidemic dynamics guided by kalman filtering techniques. Preprint.

[Prague et al., 2020] Prague, M., Wittkop, L., Clairon, Q., Dutartre, D., Thiébaut, R., and Hejblum, B. P. (2020). Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate. preprint.

Détection d'anomalies dans des séries temporelles régulières : une approche non paramétrique

Christian Derquenne

Electricité de France - Recherche et Développement - 7, boulevard Gaspard Monge - 91120 Palaiseau - christian.derquenne@edf.fr

Résumé. L'analyse du comportement des observations dans les séries temporelles est primordiale pour la prévision, la simulation, le filtrage. En effet, la présence de ruptures, d'anomalies dans une ou plusieurs séries en entrée pour prévoir une série temporelle de sortie peut entâcher la qualité de prévision. Ce problème se rencontre pour des séries temporelles régulières (températures météo, par exemple) ou irrégulières (cours de la bourse). De nombreuses méthodes de détection de ruptures et d'anomalies existent dans la littérature pour pallier ce problème. Nous proposons une approche non paramétrique de détection d'anomalies pour des séries temporelles régulières. Elle introduit une statistique robuste de test au moyen d'une distribution empirique et permet de décider du caractére anormal d'observations. Elle est appliquée sur des productions d'énergie photovoltaïque. Les résultats obtenus sont très prometteurs.

Mots-clés. Séries temporelles, détection d'anomalies, apprentissage non supervisé.

Abstract. The analysis of the behavior of observations in time series is essential for forecasting, simulation, filtering. Indeed, the presence of breaks or anomalies in one or more input series to predict an output time series can affect the quality of the forecast. This problem is encountered for regular time series (weather temperatures, for example) or irregular (stock market prices). Many methods of detecting breaks and anomalies exist in the literature to overcome this problem. We propose a non-parametric approach to anomaly detection for regular time series. It introduces a robust test statistic by means of an empirical distribution and makes it possible to decide on the abnormal character of observations. It is applied to photovoltaic energy productions. The results obtained are very promising.

Keywords. Time series, Anomalies detection, unsupervised learning.

1 Contexte - objectif

Les séries temporelles se décomposent généralement sous forme de tendance, saisonnalité, volatilité et bruits. Cependant leurs comportements peuvent être plus ou moins réguliers selon le domaine d'application. Par exemple, pour des séries généralement irrégulières, des pics peuvent être observés au sein de l'évolution des prix de l'électricité, du gaz, du pétrole, lors de situations tendues du marché, mais sur de très courtes périodes. De même des sauts en tendance, en niveau, en variabilité apparaissent dans des séries financières, comme le FTSE. A l'opposé, nous trouvons des séries temporelles dont le comportement est généralement plus régulier, comme par exemple, l'évolution de la température extérieure, la production d'énergie photovoltaïque, les courbes d'électrocardiogramme, ou dans une moindre mesure la consommation d'électricité. En effet, elles sont généralement liées à des cycles répétitifs. Cependant, quel que soit le type de séries

temporelles, leur modélisation, qu'elles soient en entrée ou en sortie, peut se révéler très délicate et doit faire appel à une grande expérience. En effet, l'apparition de comportements anormaux dans des séries régulières ou irrégulières peut biaiser les résultats d'un modèle de prévision. Tout l'art est alors d'identifier le bon grain du mauvais grain. Des critères robustes de modélisation ou des méthodes de détection d'atypicités "offline" ou "online" sont donc requis pour obtenir des prévisions ou des estimations fiables [Choudhary D., 2017]. Nous pouvons distinguer, cependant deux types de détection : la détection de ruptures et la détection d'anomalies. Le premier type consiste à rechercher des changements de comportements (pouvant être durables), en moyenne, en variance, en distribution, ..., alors que le second recherche plutôt des changements brusques de moyenne ou courte durée, tels que des pics, des creux, ... Les nombreux algorithmes développés pour la détection de ruptures sont généralement fondés sur la segmentation de signaux [Basseville et al., 1992]. Les auteurs relèvent deux ensembles d'approches : paramétriques et non paramétriques. Par exemple, des modèles de type espace-état à changements additif et dynamique permettent de détecter simplement un changement avec un retard, d'autres raffinent la position de la frontière de segments (l'instant où la rupture s'est produite). Il est aussi possible d'estimer les paramètres caractéristiques du signal avant et après la rupture. Les approches non paramètriques construisent plutôt des représentations symboliques, notamment en utilisant généralement l'apprentissage non supervisé. La détection d'anomalies repose principalement sur des approches non paramétriques en utilisant par exemple des statistiques robustes, telles que les rangs [Friend et al., 2007]. Les modèles de prévision robustes peuvent faire appel à l'algorithme LSTM (Long-Short Term Memory) conjoint à des techniques de Deep Learning [Maya et al., 2019]. L'élimination de valeurs atypiques peut être réalisée en amont pour pouvoir travailler sur des séries temporelles agrégées. Il existe également des tests non paramétriques permettant de décider de la présence ou de l'absence d'anomalies à l'aide d'approches bayesiennes, comme par exemple le modèle "Bernoulli Detector" pour des séries univariées et multivariées [Harlé, F, 2006] ou encore après segmentation de la série temporelle pour détecter des segments anormaux [Derquenne Ch., 2014]. Enfin, des réseaux bayesiens avec des graphes de dépendance entre signaux sont également mis en oeuvre [Harlé F., 2006].

Dans le cas présent, nous avons choisi de nous placer dans le cadre de la détection d'anomalies pour des séries régulières. La problématique étudiée est celle de la reconstitution de séries de production photovoltaïque réalisée par des stations. Ces séries peuvent servir à construire des modèles de prévision de consommation ou pour produire des simulations utilisées dans des chaines complexes en management de l'énergie. Nous proposons une méthode non paramétrique pour la détection d'anomalies. Elle introduit une statistique robuste de test au moyen d'une distribution empirique. Elle permet de décider du caractére anormal d'observations ou de plages d'observations. Enfin, elle peut être utilisée pour le offline et le online. La section suivante décrira pas à pas cette nouvelle approche, la suivante appliquera cette méthode sur un jeu de données mimant la production photovoltaïque. La dernière partie tirera des enseignements de cette approche et proposera des voies futures.

2 Méthode de détection d'anomalies

La démarche générale de cette approche est constituée de sept étapes : (i) découpage de la série, (ii) catégorisation des segments, (iii) agrégation des données (par exemple, passage de points 10 minutes à une heure), (iv) construction de distributions marginales des catégories, (v) introduction de(s) statistique(s) de test, (vi) établissement d'une distribution empirique pour chaque statistique de test, (vii) décision du statut "anomalie" à l'aide d'un seuil.

Soit $Y_{t,(t=1,T)}$, une série temporelle régulière périodique. Chaque Y_t est par exemple, une production photovoltaïque par point 10 minutes et T est le nombre de valeurs relevées sur plusieurs années.

(i) Cette étape de découpage de la série en M ***intervalles (catégories)*** permet de raisonner sur des grandeurs comparables (par mois, par exemple). En effet, les plages horaires d'ensoleillement sont différentes au mois de janvier et au mois de juillet. Pour cela, nous posons

$$I_0 = 0,\ I_1 =]0; \mathbf{max}_t Y_t/M]\ ...\ I_m =](m-1) \times \mathbf{max}_t Y_t/M; m \times \mathbf{max}_t Y_t/M]\ ...$$
$$...\ I_M =](M-1) \times \mathbf{max}_t Y_t/M; \mathbf{max}_t Y_t] \quad (1)$$

où m est le numéro de la catégorie.

(ii) La catégorisation des segments découle du découpage précédent et permet de raisonner sur des valeurs discrètes (catégories ordonnées), tel que : $b_t = m \times 1_{[Y_t \in I_m]}$ avec $b_t \in [0; M]$.

(iii) L'agrégation des données résume, si nécessaire, les données catégorielles (passage de points 10 minutes à des points horaires). Nous proposons la statistique de la médiane, robuste à l'égard de catégories atypiques : $b_{j,h} = \underset{t \in (j,h)}{\mathbf{med}} b_t$ calculée pour une heure h d'un jour j.

(iv) La construction de la distribution marginale des catégories permet d'obtenir les proportions observées des catégories pour chaque point horaire. Chaque proportion individuelle est de la forme :

$$f_m(h) = N_m(h) / \sum_{l=0}^{M} N_l(h) \quad (3)$$

où $N_m(h) = \sum_{j=1}^{n_h} 1_{[b_{j,h}]}$ et n_h est le nombre de jours de la période étudiée pour l'heure h.

(v) La statistique de test que nous introduisons permet de mesurer l'absence d'une anomalie pour une heure h d'un jour j donné, telle que :

$$S_{(j,h)} = \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h) \quad (4)$$

L'interprétation de (4) est la suivante. Si la catégorie $b_{j,h}$ observée pour l'heure h du jour j correspond à une fréquence faible d'apparition dans la distribution (3), cela entrainera une petite valeur de $S_{j,h}$, alors elle sera susceptible de correspondre à une anomalie.

Cette statistique peut également être utilisée pour détecter globalement si un jour j est anormal. Nous proposons trois statistiques de test. La première calcule la moyenne des statistiques horaires $S_{j,h}$, elle a l'avantage de lisser les valeurs ératiques au sein d'une période (ici, la journée), mais elle est influencée par des comportements trop atypiques. Les statistiques robustes de

la médiane et LMS (Least Median Squares introduite par Rousseeuw en 1984) permettent de pallier ce problème. L'interprétation de (5), (6) et (7) est identique à celle de (4).

$$S_j^{(moy)} = \sum_{h\in\Omega} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h)/\mathbf{card}(\Omega) \tag{5}$$

où Ω correspond aux heures durant lesquelles les productions PV sont rélevées.

$$S_j^{(med)} = \underset{h\in\Omega}{\mathbf{med}} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h) \tag{6}$$

$$S_j^{(lms)} = \underset{h\in\Omega}{\mathbf{lms}} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h) \tag{7}$$

(vi) La construction d'une distribution empirique pour chaque statistique de test permet de tenir compte de toutes les valeurs observées pour chaque jour (global) ou pour chaque heure (individuel). Nous avons choisi une approche purement non paramétrique pour établir cette distribution. En effet, celle-ci peut être réactualisée au fur et à mesure de l'arrivée des données, ce qui peut se révéler utile notamment pour une utilisation online. La distribution est notée D_h et sa fonction de répartition empirique prend la forme (8) pour la statistique $S_{j,h}$. Il est également possible de calculer cette distribution pour les trois statistiques (5), (6) et (7). Ces distributions sont asymétriques avec une queue à gauche.

$$F_h(S_{j,h}) = \mathbf{Pr}[D_h < S_{j,h}] \tag{8}$$

(vii) Le calcul de la p-valeur est effectuée à partir de la distribution précédente, construite dans un cadre d'absence d'anomalie et simuler par méthode de Monte-Carlo, pour tester si une observation est atypique avec le jeu d'hypothèses (9). Alors pour une valeur observée $s_{j,h}^{(obs)}$, nous obtenons une p-valeur issue de la distribution simulée sous H_0.

$$H_0 : Y_t \text{ est "normale" } vs\ H_1 : Y_t \text{ est une "anomalie"} \tag{9}$$

La règle de décision est la suivante : si la p-valeur associée à $s_{j,h}^{(obs)}$ est inférieure à un risque de première espèce α donné alors l'heure h du jour j est une anomalie. Cette règle de décision peut également être appliquée pour détecter si globalement un jour j est anormal à l'aide des trois statistiques (5), (6) et (7) introduites dans l'étape (v).

3 Application de l'approche de détection d'anomalies

Afin d'évaluer la qualité de l'approche proposée, nous avons simulé un jeu de données possédant les mêmes caractéristiques que des séries temporelles de production photovoltaïque. La démarche proposée en sept étapes est détaillée pas à pas sur ce jeu de données dans ce qui suit. Nous avons à notre disposition huit années observées (2011 à 2018, par exemple) relevées par point 10 minutes. La figure 1(a) montre l'évolution de la production photovoltaïque (PV) pour le mois de janvier 2011. Chaque période correspond à un jour dans la plage 7h00 à 19h00 car la production est nulle en dehors de celle-ci, la pointe se situe vers 13h00. Nous pouvons détecter que le 3

janvier qui est très perturbé en termes de forme, ainsi que le 28 janvier dont la production est très "pointée" et particulièrement, les 29 et 30 janvier qui présentent de très faibles productions. Ce dernier cas peut être rencontré lors de pannes de stations PV. La figure 1(b) affiche cette même série temporelle, mais représentée sous forme de catégories (étapes (i) et (ii)). Nous avons choisi de partager les données sur une échelle de 10 catégories. Les trois jours atypiques sont encore plus visibles et auxquels vient se rajouter le 21 janvier qui apparaissait peu (fig. 1(a)).

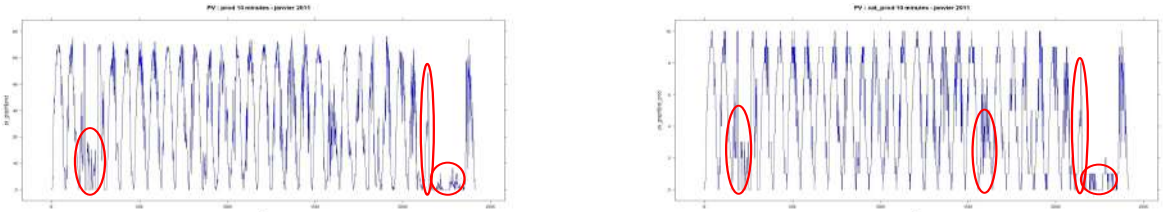


Figure 1: (a) Production PV points 10 minutes (b) Production PV catégorisées points 10 minutes

L'étape 3 d'agrégation à l'aide de la médiane des catégories pour passer des points 10 minutes à une heure (cf. fig 2(a)) confirme les résultats précédents. La figure 2(b) fournit les distributions marginales horaires sur les mois de janvier des 8 années d'étude construites à l'aide de (iv). Les heures sont représentées sur l'axe des abscisses ; les 10 catégories se situent sur l'axe des ordonnées. On distingue nettement l'évolution de la production PV tout au long d'une journée "normale" : fortes fréquences pour les catégories les plus faibles en début et fin de journée, répartitions assez étendues en milieu de matinée et d'après-midi, catégories élevées sur la plage méridienne.

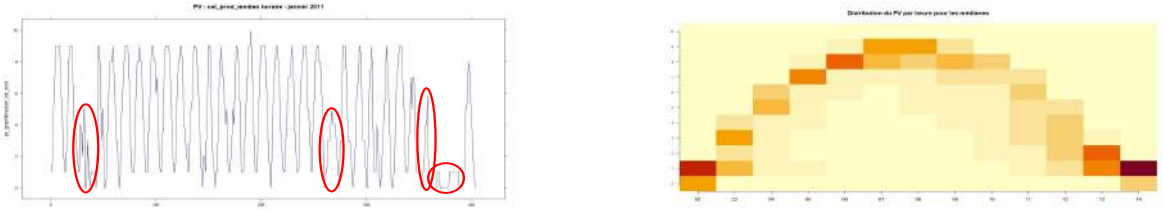


Figure 2: (a) Production PV agrégée horaires (b) Distributions marginales horaires

Tentons de répondre aux deux questions suivantes : "y a-t-il des productions PV horaires anormales le 3 janvier 2011 ?" ; "globalement la production PV de cette même journée est-elle anormale ?". Par exemple, la valeur observée de $s_{j,13}^{(obs)} = 0,0092$ et le seuil associé $F_{13}^{-1}(0,1)$ cela correspond à une p-valeur=0,045. Par conséquent, la production PV du 3 janvier 2011 à 13h00 est anormale, pour $\alpha = 0,05$ fixé (cf. étape (vii)). Les figures 3(a) et 3(b) affichent les productions PV sur lesquelles les anomalies sont représentées en rouge. La figure 3(a) correspond aux anomalies horaires sur laquelle nous pouvons en distinguer un bon nombre, les 3, 21, 28, 29 et 30 janvier, et quelques unes parsemées à d'autres dates, alors que sur la figure 3(b) relative aux anomalies journalières calculées à l'aide de la statistique (6) pour la médiane, seuls les 3, 28, 29 et 30 janvier apparaissent. Elles correspondent respectivement à des p-valeurs égales à 0,0391 ; 0,0018 ; 0,0001 ; 0,0018.

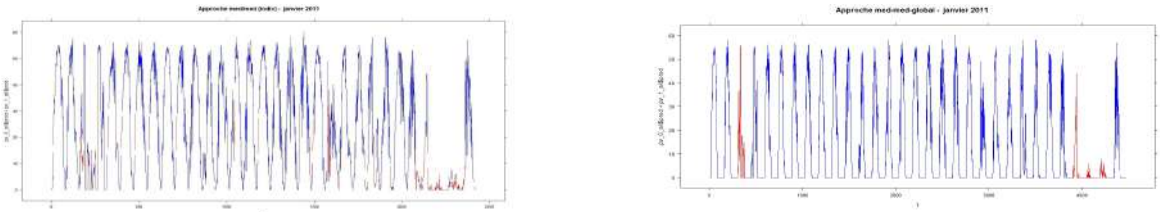


Figure 3: (a) Production PV anomalies horaires (b) Production PV anomalies journalières

4 Apports, applications et voies futures

La méthode de détection d'anomalies dans des séries temporelles régulières proposée est fondée sur une approche non paramétrique. Elle comporte une étape de simplification de la série au moyen d'un découpage des données afin d'obtenir des valeurs discrètes pour des intervalles de temps donnés (par exemple, horaires), puis une distribution pour chacun de ceux-ci est calculée. L'étape suivante consiste à construire une procédure de test pour décider si une ou ou plusieurs valeurs sont anormales. Pour cela, des statistiques de test sont introduites avec leurs distributions empiriques sous H_0 et le statut d"anormalité est décidé au moyen d'un seuil issu de la fonction de répartition empirique sous H_0 pour un niveau de risque α fixé. Cette méthode a été utilisée en offline dans le cadre de reconstitution de productions PV et a fourni des résultats très prometteurs. Les voies futures sont tout d'abord de comparer l'approche proposée à d'autres méthodes, telles que l'algorithme LSTM (Long-Short Term Memory), ou encore des tests non paramétriques fondées sur des approches bayesiennes, comme par exemple le modèle "Bernoulli Detector", ou encore des réseaux bayesiens munis de graphes de dépendance entre signaux. La seconde voie future sera d'étendre cette approche pour des stratégies online.

Bibliographie

[1] Basseville M., Nikivorov I. (1995): *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Inc.

[2] Choudhary D., Kejariwal A., Orsini F. (2017): On the Runtime-Efficacy Trade-off of Anomaly Detection, *arXiv:1710.04735v1*, MZ Inc.

[3] Derquenne Ch., (2014): Detection of similar behaviors and abnormal segments in time series, *Compstat*, Genève, Suisse.

[4] Fried R., Gather U. (2007): On rank tests for shift detection in time series, *Department of Statistics, University of Dortmund*, Germany.

[5] Harlé F., (2006): *Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance.* Thèse de doctorat, Université Grenoble-Alpes, France.

[6] Maya S., Ueno K., Nishikawa1 T. (2019): dLSTM: a new approach for anomaly detection using deep learning with delayed prediction, *International Journal of Data Science and Analytics*, **8**, 137-164.

Algorithmes de recherche dans les graphes pour l'optimisation de la maintenance prédictive de réseaux physiques : Application à la priorisation des chantiers du réseau d'assainissement de la ville de Bordeaux

C. Dumora[1] & V. Couallier[2] & C. Leclerc[3] & J. Bigot[4]

[1] *Asio, dumora.christophe@gmail.com*
[2] *Institute of Mathematics of Bordeaux, Bordeaux University, France, vincent.couallier@u-bordeaux.fr*
[3] *SUEZ-Le LyRE, France, cyril.leclerc@suez.com*
[3] *Institute of Mathematics of Bordeaux, Bordeaux University, France, jeremie.bigot@math-u-bordeaux.fr*

Résumé. Abstract. Les réseaux physiques de distribution tels que les réseaux d'eau potable, les réseaux d'assainissement, les réseaux électriques ou de télécommunication, sont des éléments des agglomération modernes (smart cities) et l'optimisation de leur maintenance dans des budgets maîtrisés est un sujet sensible. A l'aide des données stockés dans un GIS qui localisent l'ensemble des unités, et de données structurelles, ainsi que d'un historique des interventions, nous proposons une méthodologie d'optimisation des chantiers de renouvellement basés sur l'analyse du graphe du réseau par un algorithme de recherche (avec contraintes) de sous-graphes. Le résultat est une hiérarchie ordonnées de sous-graphes connectés qui peuvent directement être proposés au service de gestion patrimoniale. Le critère optimisé repose sur une agrégation spatiale d'indices de criticité, de fiabilité, de faisabilité, qui sont mesurables ou prédictibles au niveau de l'unité basique. Une application sur le réseau d'assainissement de la métropole de Bordeaux illustre la méthode.

Mots-clés. graphes, algorithme BFS, optimisation du renouvellement, réseaux d'eau

Abstract. Physical networks, such that potable water distribution networks, sewerage networks, electric networks made of power lines and transformers, telecommunication networks, are fundamental constituents of a complex urban monitored and optimized infrastructure. All these systems share the property of being computerized in a GIS (Geographic Information System). This paper proposes a methodology to provide optimized allocation of preventive maintenance budget for renewal worksites. Each potential worksite constitutes a geographical area that contains some units to renew. Each unit is identified by its localization, the indexes of units with which it is connected and some covariates, such that (for our example of a wastewater network) length, material, age, risk priority number, criticality index, probability to fail, etc.. . With the theory of graph, it is possible to construct a search algorithm that provided ranked potential worksites to

investigate. A real application on the sewerage network of Bordeaux city is provided to demonstrate the feasibility.

1 Introduction

Les réseaux physiques de distribution tels que les réseaux d'eau potable, les réseaux d'assainissement, les réseaux électriques ou de télécommunication, sont des éléments structurants des agglomération modernes (smart cities) et l'optimisation de leur maintenance dans des budgets maîtrisés et en minimisant l'impact sur la ville est un sujet sensible. La plupart de ces réseaux partagent la propriété d'être numérisés dans un SIG (système d'information géographique) et sont le plus souvent enterrés. Le renouvellement préventif (remplacement par une unité neuve identique) est alors la principale opération de maintenance des unités dégradés (avant la défaillance). Dans ce cas, choisir quand et ou intervenir est un enjeu crucial. Un autre aspect concerne la réglementation qui peut imposer certaines phases de la politique de maintenance préventive, conduisant à des problèmes d'optimisation mono ou multi-objectifs. Un de ces critères est la gêne occasionnée par les chantiers. Ceux-ci doivent dons inclure un ensemble d'unités voisines géographiquement, à changer aussi rapidement que possible, et, dans notre exemple d'application, dans une même rue.

Pour modéliser ce problème d'optimisation de maintenance, nous proposons d'utiliser la théorie des graphes, et un algorithme de recherche de sous-graphes adaptés à nos contraintes.

On considère un graphe $G(V, E)$ constitués de N_E arrêtes et N_V noeuds. Le graphe non dirigé est valué sur les arrêtes, c'est à dire que chaque arrête $e \in E$ constitue également un individu d'un jeu de données sur lequel p variables statistiques sont observées. Pour plus de détails sur les notations, on peut se référer à (Kolaczyck, 2009)

Vis à vis de notre cadre d'optimisation de la maintenance, et pour l'application visée on peut envisager disposer des variables suivantes observées sur les arrêtes du graphe : (i) des variables structurelles : longueur, diamètre, matériau, age de pose; (ii) des variables de fiabilité : indice de criticité, indice de probabilité de défaillance, ... ; (iii) des variables géographiques : rue, commune, ...

On note $\mathcal{G}$ l'ensemble de tous les sous-graphes connectés de G. Un sous-graphe $g \in \mathcal{G}$ qui contient un ensemble d'arêtes $e \in V_g, g \in \mathcal{G}$ peut donc représenter une zone géographique sur laquelle opérer un chantier de renouvellement.

2 Notation et description de l'algorithme

Un algorithme de parcours de graphe a été utilisé pour définir toutes les combinaisons de chaines simples $\mu(u, v)$ représentant des ensembles d'unités adjacentes et maximisant un

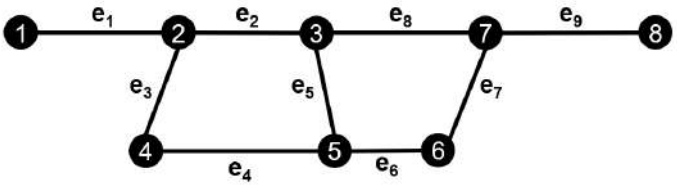


Figure 1: Un graphe planaire $G = (V, E)$ avec l'ensemble des sommets $V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ et l'ensemble des arêtes $E = \{e_1 = (1, 2), e_2 = (2, 3), e_3 = (2, 4), e_4 = (4, 5), e_5 = (3, 5), e_6 = (5, 6), e_7 = (6, 7), e_8 = (3, 7), e_9 = (7, 8)\}$

critère de priorité. Ces chaînes simples constituent ainsi des sous-graphes $g \in \mathcal{G}$.

Un parcours en largeur (ou BFS, pour Breath-First Search en anglais) est utilisé pour déterminer toutes les chaînes simples de chacun des sous graphes. Soit $g \in \mathcal{G}$ un sous-graphe de G donné, on peut alors obtenir n chaînes simples $\mu_n(u, v), \forall u, v \in V_i, u \neq v$. Etant dans un graphe pondéré on s'intéresse ici aux poids des chaînes constituées pour déterminer si celles-ci sont valides.

Les contraintes sont vérifiées durant le parcours en largeur du graphe et nous permet de déterminer toutes les chaînes simples représentant alors toutes les combinaisons d'arêtes adjacentes respectant les contraintes qualitatives ou quantitatives. Il est alors possible dans chaque sous-graphe de hiérarchiser les chaînes simples selon un critère qualitatif et ainsi déterminer la chaine simple qui maximise un critère reflétant l'objectif.

3 Application au réseau d'assainissement

La méthodologie a été appliquée pour construire un outil de recherche et d'optimisation de chantiers de renouvellement d'un réseau d'assainissement. Le module présenté est appliqué sur un réseau de graphe représentant le réseau d'assainissement de la Métropole Bordelaise.

Le réseau $G(V, E)$ est composé de $N_V = 217229$ nœuds et $N_E = 215095$ arêtes. Les arêtes représentent les canalisations du réseau d'assainissement et les nœuds tous les accessoires pouvant connecter deux canalisations entre elles, par exemple des regards. Chaque arête $E \in G(E, V)$ dispose donc de caractéristiques structurelles ou opérationnelles permettant de décrire les règles métiers utilisées lors de la constitution de chantiers de renouvellement, par exemple : (i) adresse du chantier : les travaux de renouvellement nécessitant de bloquer la voirie, ils doivent être situés dans la même rue afin de minimiser l'impact sur le réseau routier; (ii) longueur du chantier : la longueur maximale l_{max} d'un chantier de renouvellement est fixée; (iii) coût maximal du chantier : chaque chantier ne peut dépasser un certain montant fixé C_{max} par la direction de l'Eau en fonction de différents critères économiques.

On considère $\mathcal{G}$ l'ensemble des sous-graphes respectant la contrainte spatiale d'appartenance à la même rue. Pour chacun de ces sous-graphes l'algorithme BFS est utilisé pour obtenir toutes les chaînes simples respectant les contraintes de coût et longueur. Dans le cas du réseau de la métropole de Bordeaux, 1 880 037 chaînes valides sont obtenus respectant ces contraintes.

Afin de réduire le champs des possibles et orienter la sélection sur les chantiers les plus prioritaires, un premier critère intra-rue est calculé pour chacune des chaînes.

$$C_{intra} = \sum_{i=1}^{n} SP_i \times l_i \tag{1}$$

avec SP_i la note issue de PREVOIR®[1] qui traduit le niveau d'urgence de renouvellement de la canalisation i, l_i la longueur de la canalisation i et n le nombre de canalisations dans la chaîne simple. Pour chacun des sous-graphes, la chaîne simple disposant du critère intra-rue le plus élevé est sélectionnée et toutes les autres chaînes disposant d'arêtes en commun sont supprimées de la liste. On répète l'opération pour les chaînes restantes jusqu'à ce que toutes les arêtes appartiennent à une chaîne.

On dispose alors de 42056 chaînes représentant l'ensemble des chantiers qu'il est possible d'effectuer sur l'ensemble du réseau de la métropole Bordelaise.

Afin de hiérarchiser les 42 056 chantiers obtenues nous avons défini un critère inter-rue. Ce critère a pour objectif de mettre en avant les chantiers les plus à risque en normalisant le critère intra-rue par la longueur cumulée du chantier constitué. Cela permet d'éviter l'effet d'échelle et d'accorder la priorité seulement aux chantiers les plus longs. Le critère inter-rue est défini comme suit pour chacune des chaînes retenues précédemment :

$$C_{inter} = \frac{C_{intra}(j)}{L_j} \tag{2}$$

avec L_j la longueur cumulée du chantier j.

C'est à partir de ce score que l'on hiérarchise la totalité des chantiers obtenus. Le Tableau **??** présente les 10 chantiers les plus prioritaires au sens du critère inter-rue.

L'ensemble des calculs ont été réalisés sur R à l'aide de la bibliothèque *igraph*. Les recherches de chaînes simples ont été parallélisées sur chaque sous-graphe représentant l'ensemble des canalisations appartenant à la même rue afin de réduire les temps de calcul. Le code s'exécute en environ 5 heures pour le graphe de la métropole de bordeaux sur un serveur AWS disposant de 64 coeurs et 244Gb de mémoire vive.

L'algorithme fournit ainsi au gestionnaire du réseau une liste hiérarchisée des chantiers prioritaires.

[1] PREVOIR® est un outil opérationnel de SUEZ basé sur un modèle d'analyse de survie multi-états pour la prédiction des défaillances de canalisation

Rang	Nb canalisation	Longueur (m)	Coût	Critère intra	Critère inter	Score Priorité		
						min	moy	max
1	2	14.88	48 657	26.06	1.75	1.75	1.75	1.75
2	8	276.74	484 362	444.71	1.61	0	1.38	1.75
3	2	2.62	2 213	4.16	1.59	1.59	1.59	1.59
5	2	13.24	28 095	20.62	1.56	1.56	1.56	1.56
5	3	105.18	155 663	163.82	1.56	1.56	1.56	1.56
5	2	62.25	120 080	96.96	1.56	1.56	1.56	1.56
7	4	61.7	90 231	93.81	1.52	0	1.05	1.58
8	4	126.23	435 745	187.72	1.49	1.24	1.47	1.62
9	2	14.08	12 348	20.83	1.48	1.48	1.48	1.48
10	2	132.73	458 183	194.89	1.47	1.47	1.47	1.47

Table 1: Les 10 chantiers prioritaires obtenus à l'aide de l'algorithme

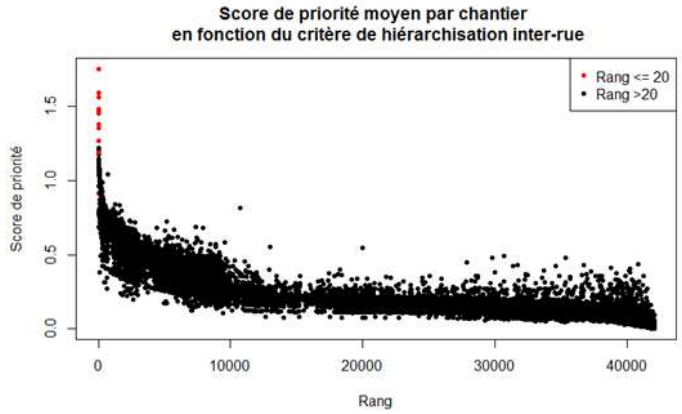


Figure 2: Scores aggrégés de priorité des 42056 chantiers éligibles

Bibliographie

Le Gat, Y. (2008), Modelling the deterioration process of drainage pipelines, *Urban Water Journal*, 5, 97–106.

Kolaczyk, Eric D. (2009), Statistical Analysis of Network Data: Methods and Models, *Springer Publishing Company, Incorporated*.

Amélioration de l'estimation d'un total en sondages par des estimateurs assistés de forêts aléatoires

Mehdi Dagdoug [1] & Camelia Goga [1] & David Haziza [2]

[1] *Université de Bourgogne Franche-Comté*
Laboratoire de Mathématiques de Besançon, Besançon, FRANCE
mohamed_mehdi.dagdoug@univ-fcomte.fr
camelia.goga@univ-fcomte.fr

[2] *University of Ottawa, Departement of Mathematics and Statistics*
Ottawa, CANADA
dhaziza@uottawa.ca

Résumé. De nos jours, les enquêtes par sondage font face à l'émergence de jeux de données complexes et de très grandes tailles. Ce type de nouvelles bases de données soulève de nouveaux défis et l'estimation de paramètres d'intérêt tels que le total, le ratio ou les quantiles basés sur des modèles paramétriques traditionnels peuvent s'avérer inefficaces. Dans ce travail, nous proposons une nouvelle classe d'estimateurs assistés par un modèle et basé sur des forêts aléatoires pour l'estimation du total d'une variable d'intérêt. Sous certaines conditions de régularité sur la variable d'intérêt, la structure des forêts et le plan de sondage, l'estimateur proposé est asymptotiquement sans biais et convergent pour l'estimation d'un total. Un estimateur convergent de la variance est suggéré et la distribution asymptotique de l'estimateur assisté par des forêts aléatoires est obtenue également permettant ainsi la construction des intervalles de confiance asymptotiques. Les simulations effectuées suggèrent que l'estimateur proposé est généralement efficace et meilleur que les estimateurs basés sur des modèles paramétriques dans le cas de relations complexes et en présence d'un très grand nombre de variables auxiliaires.

Mots-clés. Théorie des sondages, apprentissage statistique, forêts aléatoires, estimation assisté par un modèle, estimation de la variance.

Abstract. Nowadays, surveys face more and more complex data sets with a large number of variables. These new data raise many challenges and traditional parametric methods of estimation of interest parameters such as totals, ratios or quantiles may prove inefficient. In this work, we propose a new class of model-assisted estimators based on random forests. Under certain regularity conditions on the study variable, the random forest as well as the sampling design, the proposed model-assisted estimator is shown to be asymptotically design unbiased and consistent for the population total. A consistent variance estimator is proposed and the asymptotic distribution of the random-forest model-assisted estimator is obtained allowing to build confidence intervals. Simulations illustrate that the proposed estimator is efficient and can outperform state-of-the-art estimators, especially in complex and high-dimensional settings.

Keywords. Survey sampling, statistical learning, random forests, model-assisted estimation, variance estimation.

1 Introduction

Avec le développement de procédés automatiques d'acquisition de données à des échelles très fines (smart meters, objets connectés, ...), il n'est maintenant plus inhabituel de disposer de très grandes bases de données. Cette information auxiliaire très riche peut être alors utilisée pour améliorer l'estimation de type Horvitz-Thompson de paramètres d'intérêt (totaux, ratios, ...) en utilisant l'approche assistée par un modèle ("model-assisted") proposée par Särndal et al. (1992). La plupart des estimateurs de type "model-assisted" proposés dans la littérature sont basés sur un modèle paramétrique linéaire. Plus récemment, des estimateurs assistés par des modèles nonparamétriques (polynômes locaux, splines) ont été proposés en théorie des sondages mais ces estimateurs ont une efficacité réduite lorsque le nombre de variables explicatives est grand, phénomène connu sous le nom du "fléau de la dimension". Certains modèles d'apprentissage statistique récents, tels que les forêts aléatoires, (Breiman, 2001) se sont montrés très efficaces en présence de données massives. De plus, les forêts aléatoires sont capables d'utiliser efficacement à la fois variables auxiliaires qualitatives mais aussi quantitatives, ce qui est fréquent en statistique officielle, et ce sans avoir à spécifier à l'avance la forme des interactions. D'une manière générale, une forêt aléatoire est une méthode de prédiction nonparamétrique, appelée également d'ensemble, qui est basée sur une large collection d'arbres différents créés de façon aléatoire à partir des données et qui sont combinés ensuite pour obtenir une prédiction meilleure que celle obtenue avec un seul arbre.

Nous proposons dans ce travail une nouvelle classe d'estimateurs assistés par un modèle nonparamétrique et basé sur des forêts aléatoires pour l'estimation du total d'une variable d'intérêt. Nous décrivons dans la section 2 l'algorithme des forêts aléatoires considéré, puis construisons la nouvelle classe d'estimateurs assistés par des forêts aléatoires. La section 3 donne les propriétés asymptotiques de la nouvelle classe ainsi construite.

2 Estimation d'un total par assisté par des forêts aléatoires

Considérons une population finie $U = \{1, ..., k, ..., N\}$ de taille N. Nous nous intéressons à l'estimation du total de la variable d'intérêt Y au niveau de la population, $t_y = \sum_{k \in U} y_k$. Nous sélectionnons dans U un échantillon S, de taille n, avec un plan de sondage $\mathcal{P}(S)$. Les probabilités d'inclusion du premier et second ordre sont définies par $\pi_k = Pr(k \in S)$ et $\pi_{kl} = Pr(k, l \in S)$, respectivement. On suppose que $\pi_k > 0$ et $\pi_{kl} > 0$ pour tous les éléments de la population. Nous disposons de p variables auxiliaires $X_1, X_2, ..., X_p$ et nous

supposons que le vecteur $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top$ d'information auxiliaire est disponible pour chaque élément k de la population. On considère que la relation entre la variable d'intérêt et les variables auxiliaires est modélisée par le modèle de superpopulation suivant:

$$\xi : \ \mathbb{E}\left[y_k \mid \mathbf{X}_k = \mathbf{x}_k\right] = m(\mathbf{x}_k), \ k \in U,$$

où $m(\cdot)$ est une fonction inconnue. Nous proposons d'estimer le total t_y par un estimateur assisté par le modèle non-paramétrique ξ avec m estimé par des forêts aléatoires construites sur l'échantillon. Nous allons présenter d'abord les arbres et les forêts alátoires proposés dans un contexte de population infinie et adaptons ensuite leur construction au cadre de sondage.

2.1 Arbres de régression et forêts aléatoires

L'algorithme original de forêt aléatoire utilise l'algorithme CART (Breiman et al., 1984), un algorithme définissant une partition de l'espace des prédicteurs construit par des splits binaires successifs. Cet algorithme cherche la variable et la position de split pour lesquelles la différence de variance empirique dans la node avant, et après que le split ne soit réalisé, est maximisée. Plus formellement, dénotons par A la node de cardinalité $\#(A)$ considérée pour le prochain split et $\mathcal{C}_A$ l'ensemble des splits possibles dans la node A. La procédure de split est exécutée en cherchant le meilleur split, c'est-à-dire le split optimal (j^*, z^*) pour lequel le critère CART de population suivant est maximisé:

$$L_N(j,z) = \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} \left\{ (y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L} \mathbb{1}_{x_{kj} < z} - \bar{y}_{A_R} \mathbb{1}_{x_{kj} \geq z}\right)^2 \right\}, \qquad (1)$$

où $A_L = \{k \in A; x_{jk} < z\}$, $A_R = \{k \in A; x_{jk} \geqslant z\}$ et $\bar{y}_A$ est la moyenne des mesures y_k pour lesquelles les individus k appartiennent à la node A. La procédure de split continue jusqu'à ce qu'un split supplémentaire induirait la création d'une node contenant moins d'éléments qu'un nombre prédéterminé N_0. La prédiction d'un modèle de forêt aléatoire à un point $\mathbf{x}_k$ est donnée par

$$\widetilde{m}_{N,tree}(\mathbf{x}_k) = \sum_{\ell \in U} \frac{\mathbb{1}_{A_N(\mathbf{x}_k)}(\mathbf{x}_\ell)\, y_\ell}{N_N(\mathbf{x}_k)}, \qquad (2)$$

où $N_N(\mathbf{x}_k) = \sum_{\ell \in U} \mathbb{1}_{A_N(\mathbf{x}_k)}(\mathbf{x}_\ell)$ dénote la cardinalité de $A_N(\mathbf{x}_k)$, la node contenant le point $\mathbf{x}_k$.

Considérons désormais une suite $\{\theta_b^{(U)}\}_{b=1}^B$ de variables aléatoires (v.a.) indépendantes et identiquement distribuées (i.i.d.) et indépendante de $D_N = \{\mathbf{x}_k, y_k\}_{k \in U}$; celles-ci seront utilisées pour modéliser la partie aléatoire dans l'algorithme suivant. Premièrement, l'algorithme sélectionne B échantillons bootstrap, dénotés $\{D_N(\theta_b^{(U)})\}_{b=1}^B$, sans remplacement depuis D_N selon les v.a. $\{\theta_b^{(U)}\}_{b=1}^B$; chaque échantillon contient A_N couples de la

forme $(\mathbf{x}_k, y_k)$. Deuxièmement, l'algorithme estime un arbre de régression sur chaque échantillon bootstrap. Dans le contexte de forêts aléatoires, la recherche des splits optimaux à l'aide du critère (1) est réalisée en considérant seulement un sous-ensemble des p prédicteurs initialement disponibles. Ce sous-ensemble est sélectionné aléatoirement par le mécanisme aléatoire engendré par $\{\theta_b^{(U)}\}_{b=1}^B$, et peut varier au sein du même arbre d'un split à un autre. Ainsi, les B arbres de régressions définissent chacun des partitions de $\mathbb{R}^p$ potentiellement différentes. Considérons maintenant $\{\widetilde{m}_{tree}^{(1)}(\cdot, \theta_b^{(U)})\}_{b=1}^B$ une suite de B arbres de régression randomisés comme décrit ci-dessus. La prédiction d'une forêt aléatoire au point $\mathbf{x}_k$ est définie par l'expression suivante

$$\widetilde{m}_{N,rf}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{m}_{N,tree}^{(b)}(\mathbf{x}_k, \theta_b^{(U)}). \tag{3}$$

2.2 Estimation d'un total assisté par forêts aléatoires

En utilisant $\widetilde{m}_{N,rf}$, l'estimation de m sur la population, le total t_y peut être estimé par l'estimateur par différence généralisée:

$$\widehat{t}_{pgd} = \sum_{k\in U} \widetilde{m}_{N,rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widetilde{m}_{N,rf}(\mathbf{x}_k)}{\pi_k}. \tag{4}$$

Il est toutefois important de noter que l'estimateur $\widehat{t}_{pgd}$ est inutilisable en pratique car le critère (1) dépend d'observations non sélectionnées dans l'échantillon et $\widetilde{m}_{N,rf}$ est inconnu. Pour pallier ce problème, nous proposons de déterminer des partitions $\widehat{\mathcal{P}}_S = \{\widehat{\mathcal{P}}_S^{(b)}\}_{b=1}^B$ en utilisant un critère similaire au critère (1) précédent, mais en considérant une restriction aux données $D_n = \{(\mathbf{x}_k, y_k)\}_{k\in S}$. À partir de ces partitions, $\widehat{\mathcal{P}}_S$, nous obtenons une estimation de m au niveau de l'échantillon donnée par

$$\widehat{m}_{rf}(\mathbf{x}_k) = \sum_{\ell\in S} \frac{\widehat{W}_{n\ell}(\mathbf{x}_k) y_\ell}{\pi_\ell}, \qquad k \in U \tag{5}$$

où

$$\widehat{W}_{n\ell}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{A_n\left(\mathbf{x}_k, \theta_b^{(S)}\right)}(\mathbf{x}_\ell)}{\widehat{N}_n(\mathbf{x}_k, \theta_b^{(S)})}, \quad \ell \in S, \tag{6}$$

avec $\widehat{N}_n(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{\ell\in S} \pi_\ell^{-1} \psi_\ell^{(b,S)} \mathbb{1}_{A_n\left(\mathbf{x}_k, \theta_b^{(S)}\right)}(\mathbf{x}_\ell)$ et $\{\theta_b^{(S)}\}_{b=1}^B$ défini de manière similaire à $\{\theta_b^{(U)}\}_{b=1}^B$ et indépendant de S. La variable aléatoire $\psi_\ell^{(b,S)}$ indique l'appartenance de l'individu ℓ dans le sous-échantillon utilisé pour la construction de l'arbre b. Ainsi, en remplaçant $\widetilde{m}_{N,rf}(\cdot)$ avec $\widehat{m}_{rf}(\cdot)$ dans (4), nous obtenons l'estimateur de t_y assisté par une forêt aléatoire:

$$\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}. \tag{7}$$

L'estimateur $\widehat{t}_{rf}$ peut être vu comme un estimateur "bagging" dans le sens où $\widehat{t}_{rf} = (1/B)\sum_{b=1}^{B} \widehat{t}_{tree}^{(b)}$, où $\widehat{t}_{tree}^{(b)}$ dénote l'estimateur de t_y assisté par le b-ème arbre de régression et basé sur l'estimation $\widetilde{m}_{tree}^{(b)}(\cdot, \theta_b^{(S)})$ de m. On peut montrer (Dagdoug et al., 2020) que $\widehat{t}_{rf}$ peut s'écrire comme une somme pondérée de $\{y_k\}_{k\in S}$, c'est-à-dire, $\widehat{t}_{rf} = \sum_{k\in S} w_{ks} y_k$. En outre, si la forêt aléatoire sur laquelle l'estimateur en question est construit n'utilise pas de mécanisme de rééchantillonage, $\widehat{t}_{rf}$ peut aussi s'écrire sous forme de projection, i.e. $\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k)$. En pratique, les organismes nationaux de références (e.g. Insee, ...) effectuent généralement des sondages ayant pour objectif l'estimation de paramètres concernant plusieurs variables d'intérêts. Il est donc important de noter que les poids $\{w_{ks}\}_{k\in S}$ dépendent à la fois de S, de $\{y_k\}_{k\in S}$ et de $\{\mathbf{x}_k\}_{k\in U}$, ce qui n'est généralement pas le cas pour les estimateurs assistés par un modèle linéaire. Ceci implique donc que ces poids sont particulièrement adaptés à la variable qui a permis de construire la forêt, et non à une situation dans laquelle il serait requis d'estimer des paramètres en relation avec plusieurs variables d'intérêts. Une procédure de calibration par forêt aléatoire a été proposée dans Dagdoug et al. (2020) pour pouvoir répondre à cette problématique.

3 Propriétés asymptotiques

Pour obtenir les propriétés asymptotiques, nous supposons le cadre asymptotique de Isaki and Fuller (1982). Considérons pour cela une suite emboitée infinie de populations $\{U_v\}_{v\to\infty}$ de tailles $N_v \to \infty$ et d'échantillons $S_v \subset U_v$ de taille $n_v \to \infty$. Les résultats que nous décrivons dans la suite requièrent certaines hypothèses de régularité concernant le plan de sondage ainsi que la variable Y et l'algorithme de forêt sur lequel l'estimateur est construit, voir Dagdoug et al. (2020) pour plus de précisions. La plupart de ces hypothèses sont communément utilisées dans la littérature et vérifiées en pratique, voir par exemple Breidt and Opsomer (2000) et McConville and Toth (2019) pour plus de détails.

Proposition 3.1. *Il existe des constantes $C_1 > 0$ et $C_2 > 0$ telles que:*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf} - t_y\right)\right| \leqslant \frac{C_1}{\sqrt{N_v}} + \frac{C_2}{n_{0v}}, \quad \xi \text{ presque sûrement (p.s.)}, \tag{8}$$

où $\mathbb{E}_p$ représente l'espérance sous le plan de sondage.

L'équivalence suivante permet de guider notre suggestion au regard de l'estimateur de variance et de déterminer la distribution asymptotique de $\widehat{t}_{rf}$.

Proposition 3.2. *L'estimateur $\widehat{t}_{rf}$ est équivalent à l'estimateur par différence généralisée $\widehat{t}_{pgd}$:*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + o_{\mathbb{P}}(1).$$

Ce résultat nous permet de déduire la variance asymptotique de $\widehat{t}_{rf}$:

$$\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}(\pi_{kl}-\pi_k\pi_\ell)\frac{y_k-\widetilde{m}_{N,rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widetilde{m}_{N,rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{9}$$

En pratique, cette variance ne peut pas être calculée et nous proposons donc de l'estimer par

$$\widehat{\mathbb{V}}_{rf}\left(\frac{1}{N_v}\hat{t}_{rf}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v} I_k I_\ell \frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k-\widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widehat{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{10}$$

Proposition 3.3. *L'estimateur de variance $\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})$ est convergent pour $\mathbb{AV}_p\left(\widehat{t}_{rf}\right)$, c'est-à-dire,*

$$\lim_{v\to\infty}\mathbb{E}_p\left(\frac{n_v}{N_v^2}\left|\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})-\mathbb{AV}_p(\widehat{t}_{rf})\right|\right)=0.$$

Afin de pouvoir déterminer des intervalles de confiances asymptotiques, il est nécessaire de déterminer la distribution asymptotique de l'estimateur proposé qui est obtenue sous l'hypothèse supplémentaire que l'estimateur par différence généralisée $\widehat{t}_{pgd}$ suit une distribution normale.

Pour comparer les performances de $\widehat{t}_{rf}$ avec d'autres estimateurs fréquemment utilisés, nous avons effectué des simulations. Celles-ci illustrent les bonnes performances de $\widehat{t}_{rf}$ ainsi que de son estimateur de variance, y compris dans des situations particulièrement difficiles (grande dimension, faible taille d'échantillon, modèle non linéaire...).

Bibliographie

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28:1023–1053.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton.

Dagdoug, M., Goga, C., and Haziza, D. (2020). Model-assisted estimation through random forests in finite population sampling. *arXiv preprint arXiv:2002.09736*.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:49–61.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46:389–413.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Extremile Regression

Abdelaati Daouia [1] & Irène Gijbels [2] & Gilles Stupfler [3]

[1] *Université Toulouse 1 Capitole, TSE, 21 allée de Brienne, Manufacture des Tabacs, 31000 Toulouse, France,* abdelaati.daouia@tse-fr.eu

[2] *KU Leuven, Statistics and Risk Section, Celestijnenlaan, 200b - box 2400, 3001 Leuven, Belgium,* Irene.Gijbels@kuleuven.be

[3] *ENSAI & CREST, Campus de Ker Lann, 51 Rue Blaise Pascal, 35172 Bruz Cedex, France,* gilles.stupfler@ensai.fr

Résumé. Les extremiles de régression sont une version L^2 des quantiles de régression. Ils sont caractérisés par des espérances plutôt que des probabilités, et s'interprètent en tant que moyennes de minima et maxima. Ils définissent des mesures de risque cohérentes, additives comonotones, et sont des mesures de risque spectrales. On étudie ici la régression extremile en présence de covariables aléatoires. On utilise une méthode des moindres carrés localement linéaire pour l'estimation des extremiles conditionnels. On étend également l'estimation dans la queue de la distribution conditionnelle, supposée à queue lourde. On étudie l'asymptotique des estimateurs et leur comportement sur données réelles.

Mots-clés. Extrêmes, extremiles de régression, indice de queue, loi à queue lourde, moindres carrés asymétriques, quantiles de régression

Abstract. Regression extremiles define a least squares analogue of regression quantiles. They are determined by weighted expectations rather than tail probabilities. Of special interest is their intuitive meaning in terms of expected minima and maxima. In addition, they define coherent and comonotonically additive risk measures, and belong to the family of spectral risk measures. We study here extremile regression in the presence of random covariates. We rely on local linear (least squares) check function minimization for estimating conditional extremiles and deriving the asymptotic normality of their estimators. We also extend extremile regression far into the tails of heavy-tailed distributions. Extrapolated estimators are constructed and their asymptotic theory is developed. Some applications to real data are provided.

Keywords. Asymmetric least squares, Extremes, Heavy tails, Regression extremiles, Regression quantiles, Tail index

The related paper has been accepted for publication in *Journal of the American Statistical Association.*

1 Introduction

A basic tool in different scientific fields for analyzing the impact of a set of regressors X on the distribution of a response Y is quantile regression. A disadvantage of quantile

regression is that quantiles only use the information on whether an observation is below or above some specific value. However, in a financial risk management context for example, not taking into account the effective magnitude of high values of losses, might not be wise. Conditional expectiles (Newey and Powell [6]) deal with this drawback, and lead to coherent and more realistic risk measures as compared to quantile-based risk measures, as evidenced by [1] and [4], among others. An inconvenience of expectiles is their lack of transparent interpretation, due to the absence of a closed form expression. The absence of an explicit expression makes the treatment of expectiles a hard mathematical problem from the perspective of extreme value theory, for instance when it comes to estimating tail risk (Daouia *et al.* [4]).

Very recently, Daouia *et al.* [3] considered an alternative class to expectiles, called *extremiles*, which defines a new least squares analogue of quantiles. A starting point for the introduction of this class was that the unconditional τth quantile of Y, with continuous cumulative distribution function F, can alternatively be obtained from

$$q_\tau \in \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left\{J_\tau(F(Y)) \cdot \left[|Y-\theta| - |Y|\right]\right\}, \tag{1}$$

where $J_\tau(\cdot) = K'_\tau(\cdot)$, with

$$K_\tau(t) = \begin{cases} 1-(1-t)^{s(\tau)} & \text{if } \ 0<\tau\leqslant 1/2 \\ t^{r(\tau)} & \text{if } \ 1/2\leqslant\tau<1 \end{cases} \tag{2}$$

being a distribution function with support $[0,1]$, and $r(\tau) = s(1-\tau) = \log(1/2)/\log(\tau)$. See Section 2.1 in [3]. The *unconditional extremile of order* τ is then defined by substituting the absolute deviations with squared deviations, *i.e.*

$$\xi_\tau = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left\{J_\tau(F(Y)) \cdot \left[|Y-\theta|^2 - |Y|^2\right]\right\}. \tag{3}$$

Unlike expectiles, extremiles can be motivated via several angles and enjoy various interpretations and closed form expressions. For an overview on this issue, and the specific merits related to these interpretations and explicit expressions, see Daouia *et al.* [3]. In the presence of covariates, one can define conditional extremiles by considering a conditional version of (3). We pursue such an extremile regression, in a general setting.

2 Class of regression extremiles

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ be two random variables. Denote by $F(\cdot|x)$ the cumulative distribution function of Y given $X = x$ and by $q_\tau(x) = F^{-1}(\tau|x) = \inf\{y \in \mathbb{R}|F(y|x) \geqslant \tau\}$ the related conditional quantile of order $\tau \in (0,1)$. For ease of presentation, we assume throughout that $F(\cdot|x)$ is continuous. The order-τ extremile of this distribution function, as introduced in (3), defines the regression τth extremile of Y given $X = x$.

Definition 1 *Let Y given $X = x$ have a finite absolute first moment. Then, for any $\tau \in (0,1)$, the conditional order-τ extremile of Y given $X = x$ is*

$$\xi_\tau(x) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left\{J_\tau\left(F(Y|X)\right)\cdot\left[|Y-\theta|^2 - |Y|^2\right]|X=x\right\}. \tag{4}$$

Particularly useful is to look at $\xi_\tau(x)$ as the following probability-weighted moment, expected maximum or expected minimum.

Proposition 1 *Let Y given $X = x$ have a finite absolute first moment. Then, for any $\tau \in (0,1)$, we have the following equivalent closed form expressions:*

$$\begin{aligned}
\xi_\tau(x) &= \mathbb{E}\left[Y\, J_\tau\left(F(Y|X)\right)|X=x\right] = \int_0^1 J_\tau(t)\, q_t(x)\, dt = \int_0^1 q_t(x)\, dK_\tau(t), \\
\textit{and}\quad \xi_\tau(x) &= \begin{cases} \mathbb{E}\left[\max\left(Y_x^1,\ldots,Y_x^r\right)\right] & \textit{when} \quad \tau = (1/2)^{1/r} \quad \textit{with} \quad r\in\mathbb{N}\backslash\{0\}, \\ \mathbb{E}\left[\min\left(Y_x^1,\ldots,Y_x^s\right)\right] & \textit{when} \quad \tau = 1-(1/2)^{1/s} \quad \textit{with} \quad s\in\mathbb{N}\backslash\{0\}, \end{cases}
\end{aligned}$$

for independent observations Y_x^i drawn from the conditional distribution of Y given $X = x$.

3 Estimation method

Our approach is a local linear estimation based on the definition (4) which is of particular relevance when considering flexible regression specifications such as local polynomial approximations. We restrict our analysis to one-dimensional covariates X.

For a generic estimator $\widehat{F}(\cdot|x)$ of $F(\cdot|x)$, the local linear check function minimization solves the weighted least squares problem

$$\arg\min_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{i=1}^n J_\tau\left(\widehat{F}(Y_i|x)\right)\left\{Y_i - \alpha - \beta(x-X_i)\right\}^2 L\left(\frac{x-X_i}{h_n}\right) \tag{5}$$

to get the estimators $\breve{\alpha} = \breve{\xi}_{\mathrm{LL},\tau}(x)$ and $\breve{\beta} = \breve{\xi}'_{\mathrm{LL},\tau}(x)$ of $\xi_\tau(x)$ and $\xi'_\tau(x)$, respectively, where $L(\cdot)$ is a kernel function and $h_n > 0$ a bandwidth sequence. Standard weighted least squares theory leads to the following explicit solution

$$\begin{pmatrix} \breve{\alpha} \\ \breve{\beta} \end{pmatrix} = \left(\mathbf{X}_{\mathrm{LL}}^T \mathbf{W}_{\widehat{F},L} \mathbf{X}_{\mathrm{LL}}\right)^{-1} \mathbf{X}_{\mathrm{LL}}^T \mathbf{W}_{\widehat{F},L} \mathbf{Y},$$

where $\mathbf{Y}$ is the column vector of dimension n containing all Y_i, $i = 1,\ldots,n$, and $\mathbf{X}_{\mathrm{LL}}$ is the usual design matrix of the local linear fitting technique, *i.e.* the $n \times 2$ matrix with a vector of 1's as a first column, and where the second column consists of the values $x - X_i$, $i = 1,\ldots,n$. Furthermore, the weight matrix in the weighted least squares problem is

$$\mathbf{W}_{\widehat{F},L} = \mathrm{diag}\left(J_\tau\left(\widehat{F}(Y_i|x)\right) L\left(\frac{x-X_i}{h_n}\right)\right)_{i=1,\ldots,n}.$$

Clearly, the asymptotic behavior of $\widehat{F}(\cdot|x)$ will be crucial to the analysis of the asymptotic and finite-sample behavior of $\breve{\xi}_{\mathrm{LL},\tau}(x)$. One may show that, when $\widehat{F}(\cdot|x)$ has the typical rate of convergence $n^{2/5}$, the estimator $\breve{\xi}_{\mathrm{LL},\tau}(x)$ is asymptotically normal when estimating noncentral regression extremiles, namely, with $\tau \in (0, 1-1/\sqrt{2}] \cup [1/\sqrt{2}, 1)$. This should not be viewed as a restriction in practice. Indeed, by Proposition 1, regression extremiles in the right tail ($\tau \geqslant 1/2$) are most easily interpreted when the power $r(\tau) = \log(1/2)/\log(\tau)$ in (2) is an integer, since then $\xi_\tau(x) = \mathbb{E}\left[\max\left(Y_x^1, \ldots, Y_x^{r(\tau)}\right)\right]$, for independent observations Y_x^i drawn from the conditional distribution of Y given $X = x$. In this case, the condition $\tau \in [1/\sqrt{2}, 1)$ is equivalent to $r(\tau) \geqslant 2$, and hence all expected maxima and corresponding extremiles are covered by this condition, except for the conditional expectation $\xi_{1/2}(x) = \mathbb{E}(Y|X=x)$ whose estimation obviously does not require extremile regression. Likewise, regression extremiles in the left tail ($\tau \leqslant 1/2$) are interpreted as $\xi_\tau(x) = \mathbb{E}\left[\min\left(Y_x^1, \ldots, Y_x^{r(1-\tau)}\right)\right]$ when $r(1-\tau) \in \mathbb{N}\backslash\{0\}$. In this case, the condition $\tau \in (0, 1-1/\sqrt{2}]$ is equivalent to $r(1-\tau) \geqslant 2$, and so apart from $\xi_{1/2}(x) = \mathbb{E}(Y|X=x)$, all expected minima and corresponding extremiles are covered by this condition.

4 Extremal regression

In this section, we focus on *extremal regression* of a response variable $Y \in \mathbb{R}$ given a vector of covariates $X \in \mathbb{R}^d$. This translates into considering the order $\tau = \tau_n' \to 1$ or $\tau_n' \to 0$ as the sample size n goes to infinity. To ease the presentation, we restrict our extreme-value analysis to the case $\tau \to 1$. Similar considerations evidently apply to the left tail $\tau \to 0$.

4.1 Model assumption

We assume for the sake of simplicity that the response Y given $X = x$ is positive and $\mathbb{E}(Y|X=x) < \infty$. We focus on the challenging domain of attraction of heavy-tailed conditional distributions that better describe the tail structure and sparseness of the data in most applications in financial and natural sciences [2, 5, 8]. More precisely, we assume that the conditional tail quantile function $t \mapsto q_{1-t^{-1}}(x)$ is second-order regularly varying:

$$\text{(E)}\quad \forall y > 0,\ \lim_{t\to\infty} \frac{1}{A(t|x)}\left(\frac{q_{1-(ty)^{-1}}(x)}{q_{1-t^{-1}}(x)} - y^{\gamma(x)}\right) = y^{\gamma(x)}\frac{y^{\rho(x)}-1}{\rho(x)}$$

for some parameters $0 < \gamma(x) < 1$, $\rho(x) \leqslant 0$ and an auxiliary function $A(\cdot|x)$ having constant sign, with $A(t|x) \to 0$ as $t \to \infty$. We use throughout the convention that $(y^b-1)/b = \log y$ for $b = 0$, so that the right-hand side reads $y^{\gamma(x)}\log y$ if the second-order parameter $\rho(x)$ is zero. The index $\gamma(x) > 0$ tunes the tail heaviness of the conditional distribution of Y given $X = x$, with higher positive values indicating heavier conditional tails. The assumption $\gamma(x) < 1$ is tailored to our requirement that $\mathbb{E}(Y|X=x) < \infty$.

4.2 Estimation procedure and main results

Here we consider the estimation of $\xi_\tau(x)$ when $\tau = \tau'_n \uparrow 1$ at an arbitrary rate as $n \to \infty$. Under assumption (E), we have by Proposition 3 of [3], applied to the conditional distribution of Y given $X = x$, that $\xi_{\tau'_n}(x) \sim q_{\tau'_n}(x)\,\mathcal{G}(\gamma(x))$ as $n \to \infty$, where $\mathcal{G}(s) := \Gamma(1-s)\{\log 2\}^s$ and Γ is the Gamma function. This motivates the estimator

$$\widehat{\xi}^{\star}_{\tau'_n}(x) := \widehat{q}^{\star}_{\tau'_n}(x)\,\mathcal{G}\left(\widehat{\gamma}(x)\right) \tag{6}$$

obtained by substituting in suitable estimators $\widehat{q}^{\star}_{\tau'_n}(x)$ of $q_{\tau'_n}(x)$ and $\widehat{\gamma}(x)$ of $\gamma(x)$. Non-parametric local estimates of the tail quantities $q_{\tau'_n}(x)$ and $\gamma(x)$ have been proposed in the last decade by [2, 5, 8], among others. Prominent among these contributions is the Weissman quantile-type estimator

$$\widehat{q}^{\star}_{\tau'_n}(x) \equiv \widehat{q}^{\star}_{\tau'_n,\tau_n}(x) := \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\widehat{\gamma}(x)} \widehat{q}_{\tau_n}(x), \tag{7}$$

where $\widehat{\gamma}(x)$ and $\widehat{q}_{\tau_n}(x)$ are consistent estimators of $\gamma(x)$ and $q_{\tau_n}(x)$, with $\tau_n < \tau'_n$ being a tuning sequence to be selected jointly with h_n. Combining (6) and (7), we arrive at

$$\widehat{\xi}^{\star}_{\tau'_n}(x) \equiv \widehat{\xi}^{\star}_{\tau'_n,\tau_n}(x) = \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\widehat{\gamma}(x)} \widehat{q}_{\tau_n}(x)\,\mathcal{G}\left(\widehat{\gamma}(x)\right). \tag{8}$$

Here, we specialize the discussion to well-specified estimators $\widehat{q}_{\tau_n}(x)$ and $\widehat{\gamma}(x)$ in the generic form (8) of $\widehat{\xi}^{\star}_{\tau'_n}(x)$. We consider the Nadaraya-Watson type estimator $\widehat{q}_{\tau_n}(x) \equiv \widehat{F}^{-1}_{\mathrm{NW}}(\tau_n|x)$, where

$$\widehat{F}_{\mathrm{NW}}(y|x) := \sum_{i=1}^{n} \mathbb{1}(Y_i \leqslant y) L\left(\frac{x - X_i}{h_n}\right) \Bigg/ \sum_{i=1}^{n} L\left(\frac{x - X_i}{h_n}\right). \tag{9}$$

As for the choice of the conditional tail index estimator $\widehat{\gamma}(x)$, we will use in the sequel the notation $\alpha_n := 1 - \tau_n$ and $p_n := 1 - \tau'_n$, and consider the kernel estimator of [2]:

$$\widehat{\gamma}(x) = \sum_{j=1}^{J} \left[\log \widehat{q}_{1-t_j\alpha_n}(x) - \log \widehat{q}_{1-\alpha_n}(x)\right] \Bigg/ \sum_{j=1}^{J} \log(1/t_j), \tag{10}$$

where $(1 = t_1 > t_2 > \cdots > t_J > 0)$ is a decreasing list of J weights. Note that, unlike [2], we do not assume differentiability of the conditional distribution function, and therefore the distribution of Y given X is allowed to have atoms. The asymptotic normality of the corresponding regression extremile estimator

$$\widehat{\xi}^{\star}_{1-p_n}(x) := \left(\frac{\alpha_n}{p_n}\right)^{\widehat{\gamma}(x)} \widehat{q}_{1-\alpha_n}(x)\,\mathcal{G}\left(\widehat{\gamma}(x)\right)$$

follows under mild additional regularity conditions.

When choosing, for instance, the harmonic sequence $t_j = 1/j$, the variance of the limiting distribution is minimal for $J = 9$ with $V_9 \approx 1.25$.

5 Finite-sample study

The finite-sample behavior of the proposed methodology will be illustrated on two sets of real data: triceps skinfold variation data from [9] and the motorcycle insurance claims data `dataOhlsson` from the R package `insuranceData`.

Acknowledgments

The research of A. Daouia and G. Stupfler is supported by the French National Research Agency under the grant ANR-19-CE40-0013/ExtremReg project. A. Daouia acknowledges funding from the ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program). I. Gijbels gratefully acknowledges support from the Research Fund KU Leuven (projects GOA/12/014 and C16/20/002), and from Research Grant FWO G0D6619N (Flemish Science Foundation). G. Stupfler acknowledges support from an AXA Research Fund Award on "Mitigating risk in the wake of the COVID-19 pandemic".

References

[1] Bellini, F., Klar, B., Müller, A. and Gianin, E.R. (2014). Generalized quantiles as risk measures. *Insurance Math. Econom.*, **54**, 41–48.

[2] Daouia, A., Gardes, L., Girard, S. and Lekina, A. (2011). Kernel estimators of extreme level curves. *TEST*, **20**, 311–333.

[3] Daouia, A., Gijbels, I. and Stupfler, G. (2019). Extremiles: A new perspective on asymmetric least squares. *J. Amer. Statist. Assoc.*, **114**, 1366–1381.

[4] Daouia, A., Girard, S. and Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *J. R. Stat. Soc. Ser. B*, **80**, 263–292.

[5] Goegebeur, Y., Guillou, A. and Osmann, M. (2017). A local moment type estimator for an extreme quantile in regression with random covariates. *Comm. Statist. Theory Methods*, **46**, 319–343.

[6] Newey, W.K. and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819–847.

[7] Stupfler, G. (2013). A moment estimator for the conditional extreme-value index. *Electron. J. Stat.*, **7**, 2298–2343.

[8] Yu, K. and Jones, M.C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.*, **93**, 228–237.

Parameter Space Definitions for Spatial Econometric Interaction Models

Lukas Dargel

Toulouse School of Economics

lukas.dargel@tse-fr.eu

Abstract. Gravity models are traditionally used to estimate flows of goods, money or passengers between two regions. The simplest models state that the flows grow proportionally to product of the size of the regions divided by the distance. This formulation assumes independence between flows and is known to be inaccurate in the presence of spatial autocorrelation. LeSage and Pace (2008) develop a spatial econometric interaction model that accounts for spatial autocorrelation in origin-destination flows. Their model avoids the independence assumption underlying the traditional gravity model and can be estimated in a matrix formulation which makes it much more efficient than the traditional vectorized approach. However, this spatial econometric model is not defined for all possible values of the parameters and we will develop an efficient method derive a feasible parameter space for this model.

Keywords. Origin-destination flows, Cross-sectional dependence, Parameter space definition

1 Introduction

Spatial interaction models are a mathematical representation of interaction behavior between entities located at an origin and a destination. They are widely applied to explain and predict, for example, passenger-flows between airports, trade or migration flows between countries, and flows of customers from residential areas to stores. Curry (1972) showed that this type of model cannot be estimated consistently by usual ordinary least-squares, which would only be possible if the flows were independent. LeSage and Pace (2008) account for spatial dependence in gravity type models by extending the class of spatial econometric models to origin-destination flows. In general, spatial econometric models avoid the independence assumption by allowing observations to be influenced by their geographical neighbors. The neighborhood structure is reflected by spatial weight matrices which allow to compute spatial lags of the original variables. These spatial lags contain information on the average value of a variable in the neighborhood around a location and a variety of spatial econometric models use them to account for different types of spatial dependence. LeSage and Pace (2008) propose a spatial interaction model that

represents the spatial dependence structure of the flows with three neighborhood matrices. They also propose a matrix formulation of the model which leads to significant efficiency gains as it allows to perform most calculations linked to the parameter estimation with objects of dimension n (the number of spatial sites) instead of $N = n^2$ (the number of origin-destination pairs). The following section introduces this model and its matrix formulation. Section 3 explains the problem of defining the feasible parameter space for this model.

2 The model and its matrix formulation

The spatial econometric interaction model of LeSage and Pace (2008) is a spatial lag (LAG) model of order three. This means that the dependent variable y, typically a vector of origin-destination flows, is explained by some exogenous covariates Z and three autoregressive components. These autoregressive components enter the model as spatial lags of the flow vector and require to define the spatial neighborhood of all origin-destination pairs. LeSage and Pace (2008) use three matrices to define this spatial neighborhood. The destination-neighborhood matrix W_d allows to capture the mutual influence of flows that start at the same origin and end at a neighbor of the destination. Likewise, the origin-neighborhood matrix W_o represents the dependence between flows that have the same destination and start from neighboring origins. Finally, the origin-to-destination neighborhood matrix W_w represents links between origin-destination pairs whose origins and destinations are neighbors. Using these neighborhood matrices we obtain the model

$$y = \rho_d W_d y + \rho_o W_o y + \rho_w W_w y + Z\delta + \varepsilon \tag{1}$$

where the parameters δ capture the influence of the explanatory variables, ε represents a vector of random errors, and $\rho = \begin{pmatrix}\rho_d & \rho_o & \rho_w\end{pmatrix}'$ is the vector of autocorrelation parameters associated to the three possible forms of spatial dependence.

Model (1) explains the flow vector y $(N \times 1)$ with some explanatory variables and three autoregressive components. From a theoretical point of view, it is no problem to treat this model as a usual spatial econometric model with multiple weight matrices. However, this vectorized formulation is computationally inefficient because it ignores redundancies that arise from the fact that every site is at the same time origin and destination of many flows. For this reason, the spatial neighborhood matrices and the explanatory variables contain mainly duplicated information. The matrix formulation of LeSage and Pace (2008) avoids these inefficiencies by operating on the flows in their matrix representation and by deriving the elements of moments expressions such as $Z'Z$ and $Z'y$ directly from the original data. Hence, when using this matrix formulation we do not need to construct the large matrix of covariates Z $(N \times K)$ and the even larger neighborhood matrices of the origin-destination pairs W_r $(N \times N)$ (for $r = d, o, w$).

The matrix representation of the flows

To construct the flow matrix $Y(n \times n)$ we have to define an ordering of all the spatial sites at the origins and destinations of the flows. Based on this ordering, all flows are arranged into a matrix, whose columns correspond to the origins and whose rows correspond to the destinations, as is shown in (2) below.

$$\begin{pmatrix} o_1 \to d_1 & o_2 \to d_1 & \cdots & o_n \to d_1 \\ o_1 \to d_2 & o_2 \to d_2 & \cdots & o_n \to d_2 \\ \vdots & \vdots & \ddots & \vdots \\ o_1 \to d_n & o_2 \to d_n & \cdots & o_n \to d_n \end{pmatrix} \Longrightarrow \begin{pmatrix} y_1 & y_{n+1} & \cdots & y_{N-n+1} \\ y_2 & y_{n+2} & \cdots & y_{N-n+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & y_{2n} & \cdots & y_N \end{pmatrix} = Y \tag{2}$$

Using the VEC-operator that stacks the columns of a matrix (from left to right), it is then possible to derive the flow vector as $y = \text{VEC}(Y) = (y_1, y_2, \ldots, y_N)'$.

Representation of the spatial neighborhood

The ordering of the flows in (2) allows to derive the three neighborhood matrices of origin-destination pairs from a single matrix $W(n \times n)$, which describes the spatial neighborhood of the sites that correspond to the origins and destinations of the flows. The three neighborhoods of the origin-destination pairs are computed using Kronecker products $\otimes$.

$$W_d = \mathbf{I}_n \otimes W \qquad W_o = W \otimes \mathbf{I}_n \qquad W_w = W \otimes W \tag{3}$$

The above neighborhood matrices allow to compute the spatial lags of the flows efficiently from their matrix representation.

$$W_d y = \text{VEC}(WY) \qquad W_o y = \text{VEC}(YW') \qquad W_w y = \text{VEC}(WYW') \tag{4}$$

Representation of the explanatory variables

The exogenous covariates can be grouped into four sets of variables $Z = \begin{pmatrix} \iota_N & X_d & X_o & g \end{pmatrix}$, where ι_N is the constant, g is a vector of distances, and X_d and X_o are variables that describe the destinations and the origins. We can exploit the ordering in (2) to define the components of Z in terms of Kronecker products and the matrix of distances $G(n \times n)$

$$\iota_N = \iota_n \otimes \iota_n \qquad g = \text{VEC}(G) \tag{5}$$

$$X_d = \iota_n \otimes DX \qquad X_o = OX \otimes \iota_n, \tag{6}$$

where the matrices $X(k_d \times n)$ and $OX(k_o \times n)$ contain variables used to describe the destinations and the origins of the flows. Definition (5) of the explanatory variables allows to derive moments such as $Z'Z$ and $Z'y$ directly from the site attributes without having to construct the matrix Z which contains mostly redundant information.

The feasible parameter space

Unfortunately, the spatial interaction model in (1) is not defined for all values of the autoregressive parameters and we have to consider a feasible parameter space that ensures the well-behavedness of the model. The following paragraphs give an overview about the problems we encounter when defining this feasible parameter space and discuss some of the available solutions. We treat this issue first from the general perspective of a higher-order spatial LAG model. Afterwards, we develop a new solution to this problem for the special case of the spatial interaction model in (1). In general, a spatial model of order l is characterized by the fact that it uses l distinct neighborhood matrices associated to l autoregressive parameters. The spatial filter matrix of such a model is given by $A = (I - W_F)$, where $W_F = \rho_1 W_1 + ... + \rho_l W_l$.

Table 1 illustrates how the most commonly applied constraints on the parameter space for a spatial model of order one have been extended to higher order models. The first column of Table 1 defines five constraints on the feasible space for the autoregressive parameters $\rho = (\rho_1, ..., \rho_l)$ in a LAG model of order l. Columns two and three of Table 1 indicate whether a given constraint is necessary or sufficient to ensure that the model is well defined and the last column shows an expression of the feasible parameter space for the order one model. The first four constraints are ordered in decreasing restrictiveness, while the fifth constraint cannot be compared to the others because it is rather an ad hoc rule than a mathematically justified restriction.

Table 1: The feasible parameter space in spatial autoregressive models

	Constraint[a]	Necessary	Sufficient	Constraint for $(l = 1)$[b]
(I)	$\rho \in \mathbb{R}^l \setminus \mathbb{A}_1$	✓	✓	$\rho_1 \neq \lambda_i(W_1)^{-1}$, for $i = 1, ..., n$
(II)	$\rho \in \mathbb{B}_1$	✗	✓	$\lambda_{min}(W_1)^{-1} < \rho_1 < \lambda_{max}(W_1)^{-1}$
(III)	$\rho \in \mathbb{B}_2$	✗	✓	$-\phi(W_1)^{-1} < \rho_1 < \phi(W_1)^{-1}$
(IV)	$0 \leq \lvert\rho_1\rvert + ... + \lvert\rho_l\rvert < 1$	✗	✓	$-1 < \rho_1 < 1$
(V)	$-1 < \rho_1, ..., \rho_l, \sum_l \rho_l < 1$	✗	✗	$-1 < \rho_1 < 1$

[a] The set $\mathbb{A}_1$ contains all values of ρ for which at least one eigenvalue of W_F is equal to one.
[a] The set $\mathbb{B}_1$ contains all values of ρ for which all real eigenvalues of W_F are smaller than one.
[a] The set $\mathbb{B}_2$ contains all values of ρ for which all eigenvalues of W_F are smaller than one in absolute value.
[b] Let M be a square matrix, then $\lambda_i(M)$ is its ith eigenvalue, $\lambda_{min}(M)$ (respectively $\lambda_{max}(M)$) its smallest (respectively largest) real eigenvalue and $\phi(M)$ is the spectral radius of M.
[b] Constraint (IV) and (V) are typically used in conjunction with row-stochastic neighborhood matrices which always have 1 as their largest eigenvalue.

The first three constraints in Table 1 can be expressed in terms of the eigenvalues $\lambda(W_F)$ of W_F. For the order one model the eigenvalues of W_F are linked to those of W_1 by the equation $\lambda(W_F) = \rho_1 \lambda(W_1)$. This link allows to obtain $\lambda(W_F)$ for changing values of the autoregressive parameter without having to compute an eigenvalue decomposition

of the W_F matrix. In contrast, for higher order models the eigenvalues of W_F must be recomputed for every change in the autoregressive parameters, which makes it very costly to define the parameter spaces implied by the first three constraints. This difficulty explains why many empirical studies resort to the naive extensions described by constraint (IV) and (V). The main disadvantage of constraint (IV) is that is overly restrictive and constraint (V) does not allow to draw any conclusions about the consistency of the model.

We will develop a computationally efficient method that allows to define the feasible parameter space according to constraint (II) and (III), which can be done by exploiting the special structure of the neighborhood matrices in model (1).

References

Barry, Ronald Paul and Robert Kelley Pace (1999). "Monte Carlo estimates of the log determinant of large sparse matrices". In: *Linear Algebra and its Applications*, pp. 41–54.

Bivand, Roger, Jan Hauke, and Tomasz Kossowski (2013). "Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods". In: *Geographical Analysis*, pp. 150–179.

Curry, Leslie (1972). "A spatial analysis of gravity flows". In: *Regional Studies*, pp. 131–147.

Elhorst, J. Paul, Donald J. Lacombe, and Gianfranco Piras (2012). "On model specification and parameter space definitions in higher order spatial econometric models". In: *Regional Science and Urban Economics*, pp. 211–220.

Hepple, L W (1995). "Bayesian Techniques in Spatial and Network Econometrics: 2. Computational Methods and Algorithms". In: *Environment and Planning A: Economy and Space*, pp. 615–644.

Kelejian, Harry H. and Dennis P. Robinson (1995). "Spatial Correlation: A Suggested Alternative to the Autoregressive Model". In: ed. by Luc Anselin and Raymond J. G. M. Florax. Springer, pp. 75–95.

LeSage, James Paul and Robert Kelley Pace (2008). "Spatial Econometric Modeling of Origin-Destination Flows". In: *Journal of Regional Science*, pp. 941–967.

Martin, R. J. (1992). "Approximations to the determinant term in gaussian maximum likelihood estimation of some spatial models". In: *Communications in Statistics - Theory and Methods*, pp. 189–205.

Statistically Consistent Counterfactual Explanations

Lucas De Lara [1] & Alberto Gonzaléz-Sanz [2]

Institut de Mathématiques de Toulouse
[1] *lucas.de_lara@math.univ-toulouse.fr*
[2] *alberto.gonzalez_sanz@math.univ-toulouse.fr*

Résumé. Les modèles contre-factuels visent à expliquer des predictions d'algorithmes de machine learning en évaluant des questions de la forme *si l'entrée avait été différente, la sortie aurait-elle été la même ?* La littérature a principalement étudié deux méthodes opposées pour construire ces modèles : le principe du plus proche exemple contre-factuel, lequel ignore les distributions statistiques sous-jacentes, et le raisonnement causal de Pearl, qui depend d'un model inconnu en pratique. Nous étendons un travail récent qui substitue les interventions causales par un opérateur push-forward entre des distributions de probabilités. Nous montrons que cette approche conduit à des explications contre-factuelles statistiquement consistantes, sans hypothèse sur la loi des données.

Mots-clés. Contre-factuel, Loyauté

Abstract. Counterfactual models aims at explaining machine learning predictions by answering questions of the form *had the input been different, would have the output been the same?* The literature mostly focused on two divergent frameworks to build these models: the *neareast counterfactual instance* principle –which is oblivious of the latent statistical distributions, and Pearl's causal reasoning –which relies on a model unknown in practice. We extend a recent work that substituted causal interventions by a push-forward operator between probability distributions. We show that this approach leads to statistically consistent counterfactual explanations, free of prior assumptions on the data generation process.

Keywords. Counterfactual, Fairness

1 Preliminaries

1.1 Transport-based counterfactual models

A counterfactual model relates an event from a factual, observational world to an event from an alternative world. Black et al. (2020) proposed a mass transportation viewpoint, where each world is represented by a probability distribution, and where the correspondences are given by a deterministic matching. More formally, let μ_0 and μ_1 be Borel probabilities on $\mathbb{R}^d$ representing respectively the *factual* and *counterfactual* distributions.

We say that a map $T : \mathbb{R}^d \to \mathbb{R}^d$ *pushes forward* μ_0 to μ_1 if for any measurable set $B \subseteq \mathbb{R}^d$, $\mu_1(B) = \mu_0\big(T^{-1}(B)\big)$. Consider for example that μ_0 is the distribution of the women features, while μ_1 is the distribution of the men features. For any woman described by x_0, her counterfactual male counterpart is thus given by $x_1 = T(x_0)$. This approach ensures that the deformation preserves the latent distributions. Optimal transport maps (see Villani (2008) for further detail) satisfy the push-forward condition while minimizing an error between paired instances, and as such define faithful counterfactual models.

However, in practice we do not have access to μ_0 and μ_1 but to empirical measures derived from observed samples. In this work, we analyze the asymptotic behaviour of counterfactual explanations generated by the estimators of a true counterfactual model T. Set $\mathcal{X}_0$ the support of μ_0.

Definition 1 *Let T_{n_0,n_1} be an estimator of T built on a n_0-sample from μ_0 and a n_1-sample from μ_1. T_{n_0,n_1} is said to be T-admissible if*

1. *$T_{n_0,n_1} : \mathcal{X}_0 \to \mathbb{R}^d$ is continuous,*
2. *$T_{n_0,n_1}(x) \xrightarrow[n_0,n_1 \to +\infty]{a.s.} T(x)$ for μ_0-almost every x.*

We refer to our work in De Lara et al. (2021) where we build such an estimator by extending a discrete optimal transport map.

1.2 Counterfactual explanations

We address counterfactual explanations for auditing fairness in binary classifiers following the framework we presented in De Lara et al. (2021). It generalizes the one originally introduced by Black et al. (2020), restricted to data points only, to a continuous setting.

Let $h : \mathbb{R}^d \times \{0,1\} \mapsto \{0,1\}$ be a binary decision rule. The first argument of h is the feature vector x, the second one is the so-called *protected attribute* s, and by convention $h(x,s) = 1$ is the favorable outcome. Extending our gender-example, consider for example that μ_0 represents a supposedly disadvantaged population while μ_1 represents an advantaged population with respect to h.

Definition 2 *For a given binary classifier h, and a measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$, we define*

- *the FlipSet as the set of individuals whose T-counterparts are treated unequally*

$$F(h,T) = \{x \in \mathbb{R}^d \mid h(x,0) \neq h(T(x),1)\},$$

- *the positive FlipSet as the set of individuals whose T-counterparts are disadvantaged*

$$F^+(h,T) = \{x \in \mathbb{R}^d \mid h(x,0) > h(T(x),1)\},$$

- *the negative FlipSet as the set of individuals whose T-counterparts are advantaged*

$$F^-(h,T) = \{x \in \mathbb{R}^d \mid h(x,0) < h(T(x),1)\}.$$

When there is no ambiguity, we may omit the dependence on T and h in the notation.

The Flip Set characterizes a set of *counterfactual explanations* w.r.t. an intervention T. Such explanations are meant to reveal a possible bias towards the protected attribute. The partition into a positive and a negative Flip Set sharpens the analysis by controlling whether s is an advantageous attribute or not in the decision making process. As $s = 0$ represents the minority, one can think of the negative partition as the occurrences of *negative discrimination*, and the positive partition as the occurrences of *positive discrimination*. Black et al. noted that the relative sizes of the empirical positive and negative Flip Sets quantified the lack of *statistical parity*. However, the interest of such sets lies in their explanatory power rather than being proxies for determining fairness scores. By analyzing the mean behaviour of $I - T$ for points in a Flip Set, one can shed light on the features that mattered the most in the decision making process. A Transparency Report indicates which coordinates change the most, in intensity and in frequency, when applying T to a Flip Set. In what follows, for any $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ we define $\text{sign}(x) := (\text{sign}(x_1), \ldots, \text{sign}(x_d))^T$ the sign function on vectors.

Definition 3 *Let $\star$ be in $\{-,+\}$, h be a binary classifier and $T : \mathcal{X}_0 \to \mathbb{R}^d$ be measurable map. Assume that μ_0 and $T_\sharp\mu_0$ have finite first-order moments. The Transparency Report is defined by the mean difference vector*

$$\begin{aligned}\Delta^\star_{diff}(h,T) &= \mathbb{E}_{\mu_0}[X - T(X) | X \in F^\star(h,T)] \\ &= \frac{1}{\mu_0(F^\star(h,T))} \int_{F^\star(h,T)} \big(x - T(x)\big) d\mu_0(x),\end{aligned}$$

and the mean sign vector

$$\begin{aligned}\Delta^\star_{sign}(h,T) &= \mathbb{E}_{\mu_0}\big[sign(X - T(X)) \big| X \in F^\star(h,T)\big] \\ &= \frac{1}{\mu_0(F^\star(h,T))} \int_{F^\star(h,T)} sign(x - T(x)) d\mu_0(x).\end{aligned}$$

The first vector indicates how much the points moved; the second shows whether the direction of the transportation was robust. Implementing a *Flip Test* consists in computing the Transparency Reports to explain the lack of statistical parity. Rigorously, the reports must be centered-scaled to avoid inadequate explanations, as we detailed in De Lara et al. (2021).

2 Statistical Consistency

The first step for implementing the Flip Test technique is computing an estimator T_{n_0,n_1} of the chosen matching function T. The second step consists in building empirical versions of the Flip Sets and Transparency Reports for h and T_{n_0,n_1} using m data points from μ_0. The consistency problem at hand becomes two-fold: w.r.t. to m the size of the sample, and w.r.t. to the convergence of the estimator T_{n_0,n_1}. Proving this consistency is crucial, as T_{n_0,n_1} satisfies the push-forward condition at the limit only.

Consider a m-sample $\{x_i^0\}_{i=1}^m$ drawn from μ_0. We define the empirical counterparts of respectively the negative Flip Set, the positive Flip Set, the mean difference vector, the mean sign vector, and the Reference Vectors for arbitrary h and T. For any $\star \in \{-,+\}$, they are given by

$$
\begin{aligned}
F_m^\star(h,T) &:= \{x_i^0\}_{i=1}^m \cap F^\star(h,T),\\
\Delta^\star_{\text{diff},m}(h,T) &:= \frac{\sum_{i=1}^m \mathbf{1}_{F^\star(h,T)}(x_i^0)\big(x_i^0 - T(x_i^0)\big)}{|F_m^\star(h,T)|},\\
\Delta^\star_{\text{sign},m}(h,T) &:= \frac{\sum_{i=1}^m \mathbf{1}_{F^\star(h,T)}(x_i^0)\text{sign}\big(x_i^0 - T(x_i^0)\big)}{|F_m^\star(h,T)|},\\
\Delta^{\text{ref}}_{\text{diff},m}(T) &:= \frac{1}{m}\sum_{i=1}^m \big(x_i^0 - T(x_i^0)\big),\\
\Delta^{\text{ref}}_{\text{sign},m}(T) &:= \frac{1}{m}\sum_{i=1}^m \text{sign}\big(x_i^0 - T(x_i^0)\big).
\end{aligned}
$$

Note that the first four equalities correspond to the original definitions from Black et al (2020). The strong law of large numbers implies the convergence almost surely of each of these estimators, as precised in the following proposition.

Proposition 1 *Let $\star \in \{-,+\}$, h be a binary classifier, and T a measurable function. The following convergences hold*

$$
\begin{aligned}
\frac{|F_m^\star(h,T)|}{m} &\xrightarrow[m\to+\infty]{\mu_0-a.s.} \mu_0(F^\star(h,T)),\\
\Delta^\star_{diff,m}(h,T) &\xrightarrow[m\to+\infty]{\mu_0-a.s.} \Delta^\star_{diff}(h,T),\\
\Delta^\star_{sign,m}(h,T) &\xrightarrow[m\to+\infty]{\mu_0-a.s.} \Delta^\star_{sign}(h,T),\\
\Delta^{ref}_{diff,m}(T) &\xrightarrow[m\to+\infty]{\mu_0-a.s.} \Delta^{ref}_{diff}(T),\\
\Delta^{ref}_{sign,m}(T) &\xrightarrow[m\to+\infty]{\mu_0-a.s.} \Delta^{ref}_{sign}(T).
\end{aligned}
$$

In particular, theses convergences hold for an admissible estimator T_{n_0,n_1}. To address the further convergence w.r.t. n_0 and n_1, we first introduce a new definition.

Definition 4 *A binary classifier $\tilde{h} : \mathbb{R}^d \to \{0,1\}$ is separating with respect to a measure ν on $\mathbb{R}^d$ if*

1. *$H_0 := \tilde{h}^{-1}(\{0\})$ and $H_1 := \tilde{h}^{-1}(\{1\})$ are closed or open,*
2. *$\nu\big(\overline{H_0} \cap \overline{H_1}\big) = 0$.*

We argue that except in pathological cases that are not relevant in practice, machine learning always deals with such classifiers. For example, thresholded versions of continuous functions, which account for most of the machine learning classifiers (e.g. SVM, neural networks...), are separating with respect to Lebesgue continuous measures. As for a very theoretical example of non-separating classifier, one could propose the indicator of the rational numbers, which is not separating with respect to the Lebesgue measure. Working with classifiers h such that $h(\cdot,1)$ is separating w.r.t. to μ_1 fixes the regularity issues one might encounter when taking the limit in $h(T_{n_0,n_1}(\cdot),1)$.

Proposition 2 *Let $\tilde{h} : \mathbb{R}^d \to \{0,1\}$ be a separating classifier w.r.t. μ_1, and T_{n_0,n_1} a T-admissible estimator. Then, for μ_0-almost every x*

$$\tilde{h}(T_{n_0,n_1}(x)) \xrightarrow[n_0,n_1\to+\infty]{a.s.} \tilde{h}(T(x)).$$

Next, we make a technical assumption for the convergence of the Transparency Report. Let $\{e_1,\dots,e_d\}$ be the canonical basis of $\mathbb{R}^d$, and define for every $k \in \{1,\dots,d\}$ the set $\Lambda_k(T) := \{x \in \mathbb{R}^d \mid \langle x - T(x), e_k\rangle = 0\}$.

Assumption 1 *For every $k \in \{1,\dots,d\}$, $\mu_0\big(\Lambda_k(T)\big) = 0$.*

Any Lebesgue continuous measure satisfies Assumption 1. This leads to our key result.

Theorem 1 *Let $\star \in \{-,+\}$, h be a binary classifier such that $h(\cdot,1)$ is separating w.r.t. μ_1, and T_{n_0,n_1} a T-admissible estimator. The following convergences hold*

$$\begin{aligned}
\mu_0(F^\star(h,T_{n_0,n_1})) &\xrightarrow[n_0,n_1\to+\infty]{a.s.} \mu_0(F^\star(h,T)),\\
\Delta^\star_{\mathit{diff}}(h,T_{n_0,n_1}) &\xrightarrow[n_0,n_1\to+\infty]{a.s.} \Delta^\star_{\mathit{diff}}(h,T),\\
\Delta^{\mathit{ref}}_{\mathit{diff}}(T_{n_0,n_1}) &\xrightarrow[n_0,n_1\to+\infty]{a.s.} \Delta^{\mathit{ref}}_{\mathit{diff}}(T).
\end{aligned}$$

If Assumption 1 holds, then additionally

$$\begin{aligned}
\Delta^\star_{\mathit{sign}}(h,T_{n_0,n_1}) &\xrightarrow[n_0,n_1\to+\infty]{a.s.} \Delta^\star_{\mathit{sign}}(h,T),\\
\Delta^{\mathit{ref}}_{\mathit{sign}}(T_{n_0,n_1}) &\xrightarrow[n_0,n_1\to+\infty]{a.s.} \Delta^{\mathit{ref}}_{\mathit{sign}}(T).
\end{aligned}$$

As aforementioned, the assumptions of Theorem 1 are not significantly restrictive in practice.

References

Black, E. and Yeom, S. and Fredrikson, M. (2020), FlipTest: Fairness Testing via Optimal Transport, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

De Lara, L. and Gonzaléz-Sanz, A. and Loubes, J. (2021), A Consistent Extension of Discrete Optimal Transport Maps for Machine Learning Applications, *Preprint*.

Villani, C (2008). *Optimal Transport: Old and New*, Springer.

Bornes inférieures pour le compromis biais-variance

Alexis Derumigny [1] & Johannes Schmidt-Hieber [2]

[1] *Department of Applied Mathematics, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE, Delft, Netherlands. a.f.f.derumigny@tudelft.nl*
[2] *University of Twente, Drienerlolaan 5, P.O. Box 217, 7500 AE Enschede, Netherlands. a.j.schmidt-hieber@utwente.nl*

Résumé. On observe couramment que, pour les modèles non-paramétriques ou de grande dimension, les estimateurs atteignant les vitesses optimales doivent égaliser le carré de leur biais et leur variance. Bien que ce phénomène de compromis biais-variance soit très largement observé, on ne sait en fait que peu de choses sur l'existence ou non de méthodes permettant d'éviter ce compromis biais-variance. Nous proposons une stratégie générale pour obtenir des bornes inférieures sur la variance de n'importe quel estimateur dont le biais est plus petit qu'une valeur fixée. Ces résultats montrent à quel point le compromis biais-variance est inévitable, et permet de quantifier la perte de performance des méthodes qui ne le satisfont pas. Cette approche est basée sur des bornes inférieures abstraites pour la variance, impliquant des changements d'espérance par rapport à différentes mesures de probabilité ainsi que des mesures d'information comme la divergence du χ^2. Certaines des ces inégalités reposent sur un nouveau concept de matrices d'information. Dans une deuxième partie de cet article, les bornes inférieures abstraites sont appliquées à plusieurs modèles statistiques, y compris le modèle de bruit blanc Gaussien, un problème d'estimation de frontière, le modèle de séquence Gaussienne et le modèle de régression linéaire en grande dimension. Pour ces applications statistiques spécifiques, différents types de compromis biais-variance ont lieu, dont la puissance est très variable. Pour le compromis entre le biais carré intégré et la variance intégrée dans le modèle de bruit blanc Gaussien, nous proposons de combiner la stratégie générale des bornes inférieures avec une technique de réduction. Ceci nous permet de réduire le problème original à une borne inférieure sur le compromis biais-variance pour des estimateurs satisfaisant des propriétés supplémentaires de symétrie dans un modèle plus simple. Pour mettre en avant de possibles extensions de ce nouveau cadre théorique, nous étudions brièvement le compromis entre biais et déviation en valeur absolue.

Mots-clés. Décomposition biais-variance, inégalité de Cramér-Rao, statistique en grande dimension, estimation minimax, déviation en valeur absolue, estimation non-paramétrique.

Abstract. It is a common phenomenon that for high-dimensional and nonparametric statistical models, rate-optimal estimators balance squared bias and variance. Although this balancing is widely observed, little is known whether methods exist that could avoid

the trade-off between bias and variance. We propose a general strategy to obtain lower bounds on the variance of any estimator with bias smaller than a prespecified bound. This shows to which extent the bias-variance trade-off is unavoidable and allows to quantify the loss of performance for methods that do not obey it. The approach is based on a number of abstract lower bounds for the variance involving the change of expectation with respect to different probability measures as well as information measures such as the χ^2-divergence. Some of these inequalities rely on a new concept of information matrices. In a second part of the article, the abstract lower bounds are applied to several statistical models including the Gaussian white noise model, a boundary estimation problem, the Gaussian sequence model and the high-dimensional linear regression model. For these specific statistical applications, different types of bias-variance trade-offs occur that vary considerably in their strength. For the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model, we propose to combine the general strategy for lower bounds with a reduction technique. This allows us to reduce the original problem to a lower bound on the bias-variance trade-off for estimators with additional symmetry properties in a simpler statistical model. To highlight possible extensions of the proposed framework, we moreover briefly discuss the trade-off between bias and mean absolute deviation.

Keywords. Bias-variance decomposition, Cramér-Rao inequality, high-dimensional statistics, minimax estimation, mean absolute deviation, nonparametric estimation

1 Changement de mesures et matrices d'information

Pour contrôler le compromis biais-variance dans un modèle paramétrique $(P_\theta), \theta \in \Theta \subset \mathbb{R}$, on utilise souvent l'inégalité de Cramér-Rao qui affirme que tout estimateur $\widehat{\theta}$ satisfait

$$\mathbb{V}\mathrm{ar}_\theta[\widehat{\theta}] \geq \frac{\left(\partial E_\theta[\widehat{\theta}]/\partial\theta\right)^2}{F(\theta)}$$

où $F(\theta)$ désigne l'information de Fisher, en supposant certaines conditions de régularité, en particulier la dérivabilité de $E_\theta[\widehat{\theta}]$ par rapport à θ. Dans ce qui suit, nous présentons une nouvelle inégalité qui suit la même intuition, mais qui est valide sous des conditions plus faibles.

Étant données $(M+1)$ mesures de probabilité $P_0, \ldots, P_M$, nous introduisons une notion de matrice d'information qui quantifie à quel point $P_1, \ldots P_M$ représentent différentes directions autour de P_0. Supposant que P_0 domine $P_1, \ldots, P_M$, la matrice de divergence du χ^2, dénotée $\chi^2(P_0, \ldots, P_M)$ est définie comme la matrice de taille $M \times M$ dont l'entrée (j, k) est :

$$\chi^2(P_0, \ldots, P_M)_{j,k} := \int \frac{dP_j}{dP_0} dP_k - 1.$$

On remarque que la j-ème entrée diagonale de cette matrice coïncide avec la divergence $\chi^2(P_j, P_0)$ au sens usuel. Par ailleurs, la matrice de divergence du χ^2 est aussi la matrice de covariance du vecteur aléatoire $(dP_1/dP_0(X), \ldots, dP_M/dP_0(X))^\top$ sous P_0.

Théorème 1. *Pour toute variable aléatoire réelle X, on a*

$$(E_{P_0}[X] - E_{P_1}[X])^2 \leq \chi^2(P_1, P_0)\mathbb{V}\mathrm{ar}_{P_0}(X) \wedge \chi^2(P_0, P_1)\mathbb{V}\mathrm{ar}_{P_1}(X),$$

et

$$\Delta^\top \chi^2(P_0, \ldots, P_M)^{-1}\Delta \leq \mathbb{V}\mathrm{ar}_{P_0}(X),$$

où $\Delta := (E_{P_1}[X] - E_{P_0}[X], \ldots, E_{P_M}[X] - E_{P_0}[X])^\top$.

Dans le cadre d'un modèle paramétrique $(P_\theta)_{\theta\in\mathbb{R}}$, on peut réécrire la différence d'espérance pour un estimateur $\widehat{\theta}$ comme $E_{P_{\theta_0}}[\widehat{\theta}] - E_{P_{\theta_1}}[\widehat{\theta}] = Biais_{\theta_0}[\widehat{\theta}] - Biais_{\theta_1}[\widehat{\theta}] - \Delta\theta$ où $\Delta\theta := \theta_1 - \theta_0$ et $\theta_0, \theta_1 \in \mathbb{R}$. Ceci permet de relier le résultat du Théorème 1 au compromis biais-variance en choisissant $X = \widehat{\theta}$. D'autres inégalités similaires sont présentées dans l'article Derumigny et Schmidt-Hieber (2020). L'inégalité de Cramér-Rao peut alors être obtenue comme corollaire de ces résultats.

2 Estimation ponctuelle dans le modèle de bruit blanc Gaussien

Supposons qu'on observe une fonction aléatoire $Y = (Y_x)_{x\in[0,1]}$, telle que

$$dY_x = f(x)\,dx + n^{-1/2}\,dW_x,$$

où W est un mouvement brownien standard non observé. Le but est d'estimer la fonction de régression $f : [0,1] \to \mathbb{R}$ en un point donné $x_0 \in [0,1]$. Pour $R, \beta > 0$, soit $\mathscr{C}^\beta(R)$ l'ensemble des fonctions β fois continûment dérivables dont la norme de Hölder est strictement majorée par R, $\mathscr{C}^\beta(\mathbb{R}) := \mathscr{C}^\beta(+\infty)$ et

$$\gamma(R,\beta) := \sup_{K\in\mathscr{C}^\beta(\mathbb{R}):K(0)=1} \left(\|K\|_2^{-1}\left(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R}\right)_+\right)^2.$$

Remarquons que cette quantité est strictement positive si et seulement si $R > 1$. De plus, pour une constante positive C et $a \in [0, R)$, on définit

$$\overline{\gamma}(R,\beta,C,a) := \sup_{K\in\mathscr{C}^\beta(\mathbb{R}):K(0)=1} \left(\|K\|_2^{-1}\left(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R-a}\right)_+\right)^2 \exp\left(- C(R-a)^2 \frac{\|K\|_2^2}{\|K\|^2_{\mathscr{C}^\beta(\mathbb{R})}}\right).$$

De même, cette quantité est positive si et seulement si $a + 1 < R$. Nous énonçons maintenant le résultat principal sur l'estimation ponctuelle, qui est prouvé dans l'article de Derumigny et Schmidt-Hieber (2020).

Théorème 2. *Avec la convention* $(+\infty)\cdot 0 = +\infty$*, on a:*
(i): soit $\mathcal{T} = \{\widehat{f} : \sup_{f\in\mathscr{C}^\beta(R)} \left|\operatorname{Bias}_f\left(\widehat{f}(x_0)\right)\right| < 1\}$*, alors,*

$$\inf_{\widehat{f}\in\mathcal{T}} \sup_{f\in\mathscr{C}^\beta(R)} \left|\operatorname{Bias}_f\left(\widehat{f}(x_0)\right)\right|^{1/\beta} \sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \geq \frac{\gamma(R,\beta)}{n}.$$

(ii): soit $\mathcal{S}(C) := \{\widehat{f} : \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f}(x_0))| < (C/n)^{\beta/(2\beta+1)}\} \cap \mathcal{T}$*, alors, pour tout* $C > 0$*,*

$$\inf_{\widehat{f}\in\mathcal{S}(C)} \sup_{f\in\mathscr{C}^\beta(R)} \left|\operatorname{Bias}_f\left(\widehat{f}(x_0)\right)\right|^{1/\beta} \inf_{f\in\mathscr{C}^\beta(R)} \frac{\mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0))}{\overline{\gamma}(R,\beta,C,\|f\|_{\mathscr{C}^\beta})} \geq \frac{1}{n}.$$

Remarque 3. *En utilisant la décomposition biais-variance de l'erreur quadratique moyenne, pour tout estimateur* $\widehat{f} \in \mathcal{T}$*, il existe une fonction* $f \in \mathscr{C}^\beta(R)$ *qui vérifie*

$$|\operatorname{Bias}_f(\widehat{f}(x_0))|^{1/\beta}\mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) \geq \gamma(R,\beta)/n$$

et donc pour un tel choix de f*,*

$$\mathrm{MSE}_f\left(\widehat{f}(x_0)\right) = \operatorname{Bias}_f\left(\widehat{f}(x_0)\right)^2 + \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \geq \left(\frac{\gamma(R,\beta)}{n\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big)}\right)^{2\beta} + \frac{\gamma(R,\beta)}{n|\operatorname{Bias}_f\left(\widehat{f}(x_0)\right)|^{1/\beta}}.$$

Ceci montre qu'un petit biais ou une petite variance augmente toujours l'erreur quadratique moyenne. Ce résultat implique aussi que la vitesse d'estimation minimax $n^{-2\beta/(2\beta+1)}$ *peut être atteinte seulement par des estimateurs équilibrant le carré du pire biais et la pire variance.*

3 Modèle de séquence Gaussienne

Dans le modèle de séquence Gaussienne, on observe n variables aléatoires indépendantes $X_i \sim \mathcal{N}(\theta_i, 1)$. Soit $\Theta(s)$ l'ensemble de tous les vecteurs $\theta = (\theta_1, \ldots, \theta_n)$ qui sont s-sparses, c'est-à-dire avec au plus s composantes non-nulles. L'erreur quadratique moyenne d'un estimateur $\widehat{\theta}$ peut alors se décomposer comme

$$E_\theta\big[\big\|\widehat{\theta} - \theta\big\|^2\big] = \big\|E_\theta\big[\widehat{\theta}\big] - \theta\big\|^2 + \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}_\theta\big(\widehat{\theta}_i\big),$$

où le premier terme de droite joue le rôle du biais. Pour ce modèle, on sait que le risque minimax est de l'ordre de $2s\log(n/s)$, et que ce risque est atteint par un estimateur de seuillage doux, voir Donoho et al. (1992). Cet estimateur exploite la connaissance de la sparsité en rétrécissant les valeurs faibles à 0. Ce seuillage cause forcément un biais

dans l'estimation mais en même temps, il permet de réduire (davantage) la variance. En utilisant les techniques de bornes inférieures présentées dans le Chapitre 1, on peut prouver une borne inférieure pour la variance au point 0 de tout estimateur qui satisfait une borne sur le biais.

Théorème 4. *Supposons que $n \geq 4$ et $0 < s \leq \sqrt{n}/2$. Soit $\widehat{\theta}$ un estimateur de θ et γ un réel tel que $4\gamma + 1/\log(n/s^2) \leq 0.99$ et*

$$\sup_{\theta \in \Theta(s)} \left\| E_\theta\big[\widehat{\theta}\big] - \theta \right\|^2 \leq \gamma s \log\Big(\frac{n}{s^2}\Big),$$

alors, pour n suffisamment grand, on a

$$\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}_0\big(\widehat{\theta}_i\big) \geq \frac{(1-(1/2)^{0.01})}{25e\log(n/s^2)} n \Big(\frac{s^2}{n}\Big)^{4\gamma},$$

où $\mathbb{V}\mathrm{ar}_0$ est la variance de l'estimateur au point $\theta = (0, \ldots, 0)^\top$.

Comparée à l'estimation ponctuelle étudiée dans le chapitre précédent, ce résultat montre un type différent de compromis biais-variance. Ici, décroître le biais d'un facteur constant γ suffit à changer l'exposant qui apparaît dans la vitesse de la borne inférieure de la variance. Ainsi réduire excessivement la constante du biais conduit forcément à des estimateurs dont le risque est fortement sous-optimal.

4 Risque intégré dans le modèle de bruit blanc Gaussien

Dans cette section, nous montrons l'existence d'un compromis entre le biais carré intégré (IBias^2) et la variance intégrée (IVar) dans le modèle de bruit blanc Gaussien. On peut d'abord remarquer que l'erreur quadratique intégrée moyenne (*MISE*, en anglais) peut être décomposée de la façon suivante:

$$\begin{aligned} \mathrm{MISE}_f\left(\widehat{f}\right) &:= E_f\big[\big\|\widehat{f} - f\big\|^2_{L^2[0,1]}\big] = \int_0^1 \mathrm{Bias}_f^2\left(\widehat{f}(x)\right) dx + \int_0^1 \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x)\big)\, dx \\ &=: \mathrm{IBias}_f^2(\widehat{f}) + \mathrm{IVar}_f\left(\widehat{f}\right). \end{aligned} \tag{1}$$

Établir des bornes inférieures minimax pour le MISE est plus difficile que dans le cas ponctuel dans la mesure où la preuve nécessite une approche basée sur des tests multiples avec une sélection précise d'hypothèses, cf. Chapitre 2.6.1 de Tsybakov (2009).

Soit β un entier strictement positif et $S^\beta(R)$ la boule de rayon R centrée en 0 dans l'espace de Sobolev L^2 avec régularité β sur $[0,1]$, c'est-à-dire l'ensemble des fonctions de carré intégrable et dont la dérivée d'ordre β est de carré intégrable satisfaisant $\|f\|_{S^\beta([0,1])} \leq R$, où pour un domaine D, $\|f\|^2_{S^\beta(D)} := \|f\|^2_{L^2(D)} + \|f^{(\beta)}\|^2_{L^2(D)}$. Soit

$$\Gamma_\beta := \inf\Big\{ \|K\|_{S^\beta} : \|K\|_{L^2(\mathbb{R})} = 1, \operatorname{supp} K \subset [-1/2, 1/2] \Big\}.$$

Théorème 5. *On considère le modèle de bruit blanc Gaussien, avec comme espace de paramètre $S^\beta(R)$. Si $R > 2\Gamma_\beta$ et avec la convention $0 \cdot (+\infty) := +\infty$, alors on a:*

$$\inf_{\widehat{f} \in T} \sup_{f \in S^\beta(R)} \left| \mathrm{IBias}_f(\widehat{f}) \right|^{1/\beta} \sup_{f \in S^\beta(R)} \mathrm{IVar}_f\left(\widehat{f}\right) \geq \frac{1}{8n},$$

où $T := \{\widehat{f} : \sup_{f \in S^\beta(R)} \mathrm{IBias}_f^2(\widehat{f}) < 2^{-\beta}\}$.

Comme dans le cadre de l'estimation ponctuelle, les estimateurs qui ont un biais important ne sont pas très intéressants dans la mesure où ils correspondent à des procédures dont le risque au sens du MISE est sous-optimal. En utilisant la décomposition biais-variance (1), on obtient que tout estimateur $\widehat{f} \in T$ satisfait la borne inférieure:

$$\sup_{f \in S^\beta(R)} \mathrm{MISE}_f\left(\widehat{f}\right) \geq \left(\frac{1}{8n \sup_{f \in S^\beta(R)} \mathrm{IVar}_f(\widehat{f})} \right)^{2\beta} \vee \frac{1}{8n \sup_{f \in S^\beta(R)} |\mathrm{IBias}_f(\widehat{f})|^{1/\beta}}.$$

Ainsi, les estimateurs dont le biais est trop petit ou la variance trop petite auront automatiquement avoir un MISE grand. Ceci permet de prouver l'existence d'une borne inférieure pour le compromis biais-variance, correspondant à la célèbre "courbe en U" observée de façon générale en estimation non-paramétrique ainsi que dans les modèles de grande dimension. En particulier, un corollaire de ce résultat donne directement le fait que $n^{-2\beta/(2\beta+1)}$ est une borne inférieure pour la vitesse minimax du risque quadratique intégré (MISE). De plus, cette vitesse ne peut être atteinte que si le pire biais carré et la pire variance intégrée sont équilibrés de telle façon à être du même ordre.

Bibliographie

Derumigny, A. et Schmidt-Hieber, J. (2020). On lower bounds for the bias-variance trade-off, *ArXiv preprint*, arXiv:2006.00278.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., et Stern, A. S. (1992). Maximum entropy and the nearly black object, *Journal of the Royal Statistical Society Serie B*, 54, pp. 41–81.

Tsybakov, A. B. (2009), Introduction to nonparametric estimation, Springer.

Sparse Subspace K-means

Abdoul Wahab Diallo [1] & Ndèye Niang [2] & Mory Ouattara

[1] *abdoul.w.m.diallo@aims-sénégal.org*
[2] *n-deye.niang_keita@cnam.fr*
[3] *ouattaramory.sfa@univ-na.ci*

Résumé. Dans ce papier nous abordons les méthodes de classification grande dimension, plus précisément lorsque les individus sont décrits par des sous-espaces de variables. Nous proposons une nouvelle approche de sparse subspace clustering appelée Sparse Subspace K-means (SSKM) qui est basée sur une modification de la fonction de coût d'une version sparse de l'algorithme K-means. La méthode proposée est illustrée sur des données simulées et sur un jeu de données réelles. Dans sa comparaison avec les méthodes de la littérature, SSKM se montre aussi bonne ou meilleure tant au niveau des indices de qualité de partition que de la détection de variables pertinentes.

Abstract. In this paper we discuss clustering methods adapted to the case of the high dimensional data, more precisely when individuals are described by subspaces of variables. We propose a new sparse subspace clustering approach called Sparse Subspace K-means (SSKM) based on a modification of the cost function of a sparse version of K-means algorithm. The proposed method is illustrated on simulated data and a real data set. In comparison with the literature methods, SSKM has good or better performances both in terms of partition quality indices and detection of relevant variables.
Keywords. k-means, Sparse, subspace clustering, high dimensional data

1 Introduction

Nous nous intéressons au problème de la classification automatique d'individus décrits par un grand ensemble de variables. Dans ce cas de grande dimension les classes sont très souvent décrites par des sous-espaces de variables rendant ainsi de nombreuses variables non pertinentes pour la classification. Ces variables peuvent alors pénaliser l'apprentissage des algorithmes classiques de classification en masquant les classes [4] ; il devient ainsi difficile pour ces algorithmes de retourner la bonne partition.
Pour pallier à l'inefficacité des méthodes classiques, les algorithmes Subspace Clustering ont été proposés. L'objectif des méthodes de Subspace Clustering est de retrouver des classes caractérisées par des sous-espaces de variables plutôt que par l'ensemble des variables. Il existe deux principaux types d'algorithmes de Subspace Clustering : les algorithmes 'Hard Subspace Clustering' qui déterminent les sous-espaces exacts où

se trouvent les classes et les algorithmes 'Soft Subspace Clustering' qui attribuent des poids aux variables et découvrent les classes à partir de sous-espaces de variables ayant des poids élevés [2]. Ainsi dans [3] les auteurs proposent la méthode Entropy Weighting K-Means (EWKM) basée sur la détermination d'un ensemble de poids introduits dans la fonction de coût de l'algorithme des K-Means. Dans la méthode EWKM, les auteurs minimisent simultanément, l'inertie intra-classe et maximise un terme d'entropie négatif dans le processus d'apprentissage. EWKM calcule pour chaque variable des poids inversement proportionnels à leur variance dans chaque classe. Le sous-espace de variables pertinentes pour chaque classe est défini en se basant sur ces poids, facilitant ainsi l'interprétation des classes. Cependant, lorsque le nombre de variables est très grand, certains algorithmes de Subspace Clustering notamment l'algorithme EWKM [3] perdent cette propriété de faciliter l'interprétation. En s'inspirant des méthodes sparse en régression, dans Sparse K-means les auteurs ont proposé une méthode qui permet de déterminer les variables pertinentes pour la partition obtenue par la méthode des K-means [1]. Cependant contrairement aux méthodes de Subspace clustering dans lesquelles les poids sont associés spécifiquement aux classes, dans la méthode Sparse K-means la pertinence des variables est relative à la partition globalement et non aux classes individuellement.
Nous proposons d'étendre la méthode Sparse K-means au cas du Soft Subspace Clustering à travers une nouvelle méthode appelée SSKM (Sparse Subspace K-means). Cette méthode permet de mettre à zéro les poids des variables non pertinentes pour chaque classe tout en affectant des poids importants aux variables caractéristiques de la classe.
En section 2 nous présentons la nouvelle méthode SSKM et les différentes étapes de l'algorithme associé. Dans la section 3, nous faisons une comparaison de la méthode SSKM avec les méthodes EWKM [3] et Sparse K-means [1] à travers une application sur des données simulées et sur des données réelles.

2 Sparse Subspace K-means (SSKM)

Nous disposons de n individus décrits par p variables $x^j (j = 1, \ldots, p)$. x_i^j désigne la valeur de la variable pour l'individu i, K le nombre de classes et w le vecteur des poids des variables. La méthode SSKM repose sur le critère ci-dessous proposé dans [1] pour la méthode Sparse K-means :

$$\max_{C_1,\ldots,C_k,w} \left\{ \sum_{j=1}^{p} w_j \Big(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_i, x_i' \in C_k} d(x_i^j, x_{i'}^j) \Big) \right\} \tag{1}$$

sous les contraintes suivantes $||w||_2 \leq 1$, $||w||_1 \leq s$, $w_j \geq 0$, $\forall j$ avec $d(x_i^j, x_{i'}^j)$ désignant la distance euclidienne entre x_i^j et $x_{i'}^j$.
L'optimisation de la fonction objectif de Sparse K-means [1] fournit les poids $w_j (j = 1, ..., p)$ sparse pour la partition. Nous proposons une modification de la fonction objectif

(1) en ramenant la sparsité des poids au niveau des classes de la partition.
Plus précisément, le critère d'optimisation de SSKM est obtenu en modifiant (1) pour tenir compte de la pertinence de la variable dans la détermination de la classe à travers notamment le remplacement de w_j par w_j^k
On obtient ainsi la fonction objectif suivante :

$$\underset{c_1,...,c_K,w^k}{max} \left\{ \sum_{j=1}^{p} \sum_{k=1}^{K} w_j^k \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \frac{1}{n_k} \sum_{x_i, x_{i'} \in c_k} d(x_i^j, x_{i'}^j) \right) \right\} \tag{2}$$

sous les contraintes :

$\| w^k \|_2 = \sqrt{\sum_{j=1}^{p} (w_j^k)^2} \leq 1$, $\| w^k \|_1 = \sum_{j=1}^{p} |w_j^k| \leq s$ et $w_j^k \geq 0, \forall j$ et k. w^k étant le vecteur poids des variables dans la classe k. Le premier terme de (2) est lié à l'inertie totale et le second terme à l'inertie intra-classe. Ainsi, maximiser (2) revient à minimiser l'inertie intra-classe. De manière identique à Sparse K-means, le processus d'optimisation des poids des variables dans (2) se fera en utilisant la formule suivante :

$$w^k = \frac{S((a^k)_+, \Delta)}{||S((a^k)_+, \Delta)||_2} \tag{3}$$

où $a_j^k = \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \frac{1}{n_k} \sum_{x_i, x_{i'} \in c_k} d(x_i^j, x_{i'}^j)$, $(x)_+$ désigne la partie positive de x et $\Delta = 0$ si cela résulte à $||w^k||_1 < s$, autrement, $\Delta > 0$ est choisi de manière à ce que $||w^k||_1 = s$. S est défini par $S(x, c) = sign(x)(|x| - c)_+$. L'algorithme itératif pour la maximisation de (2) est présenté dans la sous section suivante.

Algorithme Sparse Subspace K-means (SSKM)

1. Initialiser w^k avec $w_1^k = ... = w_p^k = \frac{1}{\sqrt{p}}$, $1 \leq k \leq K$.
2. Itérer jusqu'à la convergence :
   — Fixer w^k et optimiser (2) par rapport à $C_1, ..., C_K$:

$$\underset{C_1,...,C_K}{min} \left\{ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_i, x_{i'} \in c_k} \sum_{j=1}^{p} w_j^k d(x_i^j, x_{i'}^j) \right\} \tag{4}$$

   en appliquant l'algorithme standard K-means aux données pondérées.
   — Fixer $C_1, ..., C_K$ et optimiser (2) par rapport à $w^1, ..., w^k$ en appliquant la formule (3).

L'algorithme SSKM repose sur un choix adéquat du paramètre de sparsité s. La sélection de ce paramètre s se fait de la même manière que pour la méthode sparse K-means [1] sur la base de permutations des données, la statistique du *Gap* et la fonction objectif de SSKM.

3 Applications

3.1 Données

Deux jeux de données nommés X_1 et X_2 ont été simulés. Les deux jeux de données contiennent 200 observations décrites par 60 variables. Les observations sont reparties en quatre classes de 50 observations chacune. Le jeu de données X_1 a été simulé de telle sorte que toutes les classes soient bien séparées et décrites par l'ensemble des variables. Par contre, pour le jeu de données X_2 chaque classe est décrite par un sous-espace de l'ensemble des variables. Les classes c_1,c_2,c_3, et c_4 sont décrites respectivement par les groupes de variables 1 *à* 15, 16 *à* 30, 31 *à* 45 et 46 *à* 60. Naturellement pour chacune des classes les variables décrivant les autres classes sont des variables de bruit.
Par ailleurs, on utilise également la base de données Multiple Features (Dutch utility maps, DMU) qui contient 2000 chiffres manuscrits regroupés en 10 classes $c_k (k = 0, \ldots, 9)$, chacune ayant 200 observations. Chaque chiffre est décrit par 649 variables qui sont réparties dans les six groupes de variables suivants :
$G1 = \{A\_i(i = 1, \ldots, 75)\}$, $G2 = \{B\_i(i = 1, \ldots, 215)\}$, $G3 = \{C\_i(i = 1, \ldots, 63)\}$, $G4 = \{D\_i(i = 1, \ldots, 239)\}$, $G5 = \{E\_i(i = 1, \ldots, 46)\}$ et $G6 = \{F\_i(i = 1, \ldots, 5)\}$.

3.2 Comparaison des performances

La table 1 contient les performances moyennes (et écart type) pour 30 répétitions des algorithmes K-means, EWKM, Sparse K-means et SSKM sur les jeux de données X_1, X_2 et DMU en terme de NMI (Normalized Mutual Information) et de RA (Indice de Rand Ajusté).

Données	Indices	K-means	EWKM	Sparse K-means	SSKM
X_1	NMI	0.95 (0.1)	0.52 (0.1)	**1** (0)	**1** (0)
	RA	0.93 (0.13)	0.45 (0.21)	**1** (0)	**1** (0)
X_2	NMI	0.76 (0.08)	0.47 (0.17)	0.89 (0)	**0.92** (0.02)
	RA	0.77 (0.12)	0.45 (0.17)	**0.92** (0)	**0.92** (0.03)
DMU	NMI	0.80	0.64	0.81	**0.83**
	RA	0.74	0.54	0.78	**0.80**

TABLE 1: Performances des classifications de K-means, EWKM, Sparse K-means et SSKM sur les jeux de données X_1, X_2 et DMU

Les comparaisons des résultats montrent que les algorithmes SSKM et Sparse K-means donnent sur X_1 des performances meilleures que les algorithmes Kmeans et surtout EWKM. En effet comme cela a été présenté dans la section (3.1), les classes sont parfaitement séparées sur le jeu de données X_1 et sans structure de type 'subspace' ; ce qui

explique la très bonne performance des algorithmes SSKM, Sparse K-means et Kmeans. En ce qui concerne le jeu de données X_2, SSKM a de meilleures performances que les autres méthodes. Cela est dû au fait que Sparse K-means considère que toutes les variables sélectionnées sont pertinentes pour toutes les classes or ce n'est pas le cas sur le jeu de données X_2 qui a une structure de type 'Subspace'. On observe également de meilleures performances de SSKM pour les données DMU.

3.3 Évaluation de la pertinence des variables

Données simulées : Les figures ci-dessous représentent les heatmaps des poids des variables dans les quatre classes du jeu de données X_2 retournés par les algorithmes SSKM, EWKM et Sparse K-means.

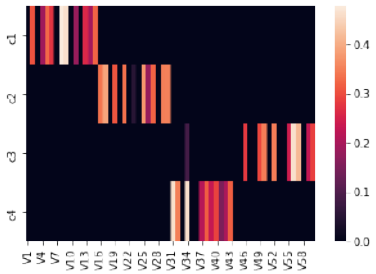


FIGURE 1: SSKM

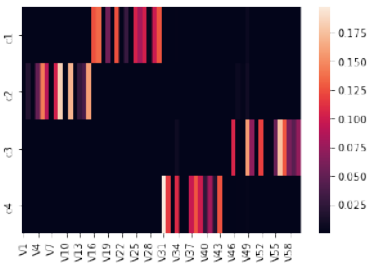


FIGURE 2: EWKM

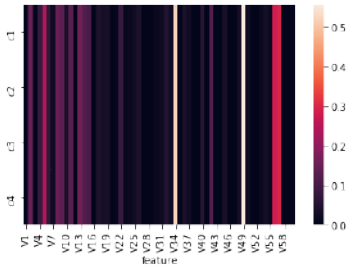


FIGURE 3: Sparse KM

On voit à travers les quatre blocs les plus coloriés de la figure (1) et (2) que les poids les plus élevés attribués par SSKM et EWKM sur les variables sont effectivement pour les variables définissant les quatre classes du jeu de données X_2 comme cela a été présenté dans la section (3.1). On observe aussi que les poids de la majorité des variables de bruit dans les classes sont caractérisés par les blocs en noir montrant ainsi l'impertinence de ces variables.
La figure (3) représente les poids affectés par Sparse K-means aux différentes variables de X_2. On observe que la grande majorité des poids des variables décrivant au départ les classes c_2, c_3 et c_4 sont soit proches de zéro ou égales à zéro. Ceci peut expliquer les faibles performance de Sparse K-means.On souligne que les deux variables ($V34$ et $V49$) ayant les poids les plus élevés sont des variables décrivant au départ respectivement les classes c_3 et c_4.

Données DMU : sur les heatmaps des poids des variables dans les dix classes du jeu de données DMU, que nous ne présentons pas ici pour des raisons de place, on observe que pour SSKM à l'exception de c_1 et c_5 les classes sont plus décrites par les variables des groupes $G4$ et les poids des variables des groupes $G1$, $G2$, $G3$, $G5$ et $G6$ dans ces classes sont très faibles. Quant à la classe c_1, elle est plus décrite par les variables des groupes

$G1$ et $G2$.
Pour l'algorithme EWKM on observe qu'une bonne partie des poids des variables des différents groupes ont des poids très faibles.
Pour la méthode Sparse K means on observe que les classes sont plus décrites par les variables des groupes $G1$, $G2$ et $G4$ contrairement aux variables des groupes $G3$ et $G5$ qui sont en majorité impertinentes pour les classes.
On peut donc conclure que la méthode proposée (SSKM) est plus performante au niveau des indices de qualité de la partition obtenue. Par ailleurs, SSKM détermine des sous-ensembles de variables pertinentes pour les classes. Plus précisément, elle permet de retenir un faible ou relativement faible pourcentage (entre 7% et 13%) des variables considérées comme pertinentes tout en gardant de bonnes performances facilitant ainsi l'interprétation.

4 Conclusion

Dans ce papier, nous avons présenté la méthode Sparse Subspace K-means (SSKM) qui est une extension de la méthode Sparse K-means [1]. Cette nouvelle méthode permet de sélectionner les variables pertinentes pour chaque classe en mettant à zéro les poids des variables non pertinentes. La méthode SSKM a été ensuite appliquée sur des données simulées et sur des données réelles. Les résultats obtenus sont satisfaisants en terme d'indice d'évaluation de la qualité de la classification et la méthode parvient à sélectionner parfaitement les variables pertinentes pour les classes.
Dans certains applications, les variables peuvent être préalablement structurées en blocs comme dans le cas des données DMU. Dans nos travaux futurs, nous envisageons d'étendre la méthode SSKM en nous inspirant des méthodes telles que FGKM [2] pour avoir une sparsité sur les blocs de variables.

Références

[1] Daniela M. Witten and Robert Tibshirani. *A framework for feature selection in clustering*. Journal of the American Statistical Association, 713–726, 2010.

[2] X Chen and Y Ye and X Xu and J. Z Huang. *A feature group weighting method for subspace clustering of high-dimensional data*. Pattern Recognition, 434-446, 2012.

[3] L. Jing and M. Ng and J. Huang. *An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data*. Knowledge and Data Engineering, IEEE Transactions, 1026-1041, 2007.

[4] L. Parsons and E. Haque and H. Lui *Subspace clustering for high dimensional data : a review*. SIGKDD Explor. Newsl, 90-105, 2004.

Étude complète de données neuronales en grande dimension, à l'aide de processus de Hawkes

Anna Bonnet [1], Charlotte Dion-Blanc [1], Sarah Lemler [2]

[1] *LPSM, UMR 8001, Sorbonne Université*
[2] *Laboratoire MICS, École CentraleSupélec, Université Paris-Saclay*

Résumé. Nous proposons d'étudier une base de données constituée de mesures effectuées sur un grand nombre de neurones au niveau de la colonne vertébrale d'une tortue. Le premier enjeu de cette étude est de détecter un réseau de communication entre les neurones à l'aide d'une modélisation des trains de *spikes* (temps d'impulsions ou de décharges) par des processus de Hawkes. Le second enjeu est d'utiliser cette information pour expliquer le comportement du potentiel de membrane d'un neurone fixé entre deux impulsions. Pour cette seconde phase nous étudions une modélisation par processus de diffusion avec des sauts modélisés par un processus de Hawkes. Cette modélisation permet d'utiliser toute la richesse des données dont nous disposons (intra et extra-cellulaires), et de représenter le lien entre les signaux reçus par un neurone et son potentiel électrique.
Mots-clés. Neurosciences, Processus de Hawkes, Inférence de graphe.

Abstract. We propose to study a neuronal dataset recorded from a large sample of neurons of the lumbar spinal cord of a turtle. The first thing at stake is to infer the connectivity graph between neurons using Hawkes processes to model spike trains. The second purpose is to use this information to explain the behavior of the membrane potential between two spikes. To that aim we use a jump diffusion model with jumps driven by a multivariate Hawkes process. This model uses the wealth of the available data (intracellular and extracellular recordings) and allows us to link the variation of the membrane potential of a central neuron to the signals received from the neurons surrounding it.
Keywords. Neuroscience, Hawkes processes, Graph inference.

1 Contexte et objectifs

En neurosciences, nous disposons de deux types de signaux : un signal extra-cellulaire qui correspond aux potentiels d'actions, *spikes* en anglais, qui peuvent être mesurés simultanément sur plusieurs neurones et un signal intracellulaire qui correspond au potentiel de membrane d'un neurone mesuré en temps continu. Nous allons travailler sur ces deux

types de données. D'une part, les spikes donnent une indication sur la façon dont le système neuronal traite une information reçue; d'autre part, l'enregistrement du potentiel de membrane peut permettre de comprendre les mécanismes internes d'un neurone. Nous disposons de toutes ces données mesurées sur un même intervalle de temps : le potentiel de membrane d'un neurone fixé et les temps de spikes de plusieurs neurones autour. L'objectif est d'abord de retrouver le graphe de connectivité de ce réseau de neurones et ensuite de modéliser le potentiel de membrane du neurone fixé en prenant en compte les signaux envoyés par les neurones avoisinants.

Nous considérons donc un réseau de neurones avec l'ensemble des neurones, dont le neurone central, pour lequel nous disposons du signal continu avec la mesure du potentiel de membrane. A partir de cette mesure nous en déduisons les temps de spikes de ce neurone. Puis, nous intégrons ce train de spikes à ceux des neurones avoisinants. Ainsi, nous pouvons estimer un réseau de communication entre ces neurones en utilisant les processus de Hawkes multivariés. Par la suite, à l'aide de la trace du potentiel de membrane du neurone central nous validons les connexions trouvées.

Enfin nous nous concentrons sur le potentiel de membrane du neurone central. Nous utilisons une modélisation originale avec un processus de diffusion avec des sauts générés par un processus de Hawkes, étudiée dans Dion & Lemler (2019) et Amorino *et al.*(2020). Bien que très simple par rapport à la complexité du phénomène observé, ce processus répond bien à l'étude de ces données, et permet de modéliser les phases entre deux spikes en prenant en compte l'influence du réseau sur les actions du neurone central.

Cette étude utilise des méthodes d'estimation récentes, adaptées aux données neuronales. La procédure complète dépend uniquement des données.

2 Données

Grâce au Pr. Rune Berg de l'Université de Copenhague (département de neuroscience), nous disposons des deux types de données issues de mesures de circuits lombaires de la colonne vertébrale d'une tortue. Elles sont étudiées par exemple dans l'article Radosevic *et al.*(2019). Nous avons 10 échantillons de trains de spikes de 249 neurones et 10 enregistrements de potentiel de membrane du neurone central, le tout sur un intervalle de 40 secondes. Une stimulation mécanique est faite au bout de 10 secondes. Nous extrayons les temps de spikes (signal discret) du neurone central à partir du potentiel de membrane (signal continu) et nous obtenons ainsi un réseau de 250 trains de spikes.

Afin d'utiliser une modélisation par processus ponctuels et des techniques d'estimation adéquates, nous devons nous assurer de la stationnarité du processus observé. Pour cela nous étudions l'outil visuel appelé PSTH (Peristimulus time histogram) à l'aide du package R `STAR`, qui donne une mesure du taux de saut du processus en fonction du temps à partir du stimulus. Après avoir sélectionné une phase satisfaisante nous gardons finalement un sous-ensemble de neurones et sur un intervalle de temps réduit.

3 Inférence

Une étape importante consiste à retrouver le graphe de connexion entre les neurones. Nous cherchons à reconstruire une matrice d'adjacence, où chaque entrée (i, j) représente l'influence du neurone j sur le neurone i, qui peut être positive (phénomène d'excitation) ou négative (inhibition).

Un outil classique pour étudier ces interactions est le "cross-correlogram" (CCH), qui, pour chaque paire de neurones, donne une estimation de la probabilité de saut du premier neurone autour des sauts du second neurone. La limite de cet outil est qu'il permet uniquement d'étudier les interactions deux à deux, et peut par exemple détecter des corrélations erronées entre deux neurones qui interagissent avec un même neurone commun. Nous proposons donc d'utiliser un modèle multidimensionnel et deux méthodes d'estimation associées, que nous présentons ci-dessous.

Nous choisissons de modéliser les trains de spikes par un processus de Hawkes multivarié noté $N_t = (N_t^{(1)}, \dots, N_t^{(M)})$ (chaque coordonnées $N_t^{(i)}$ est le processus de comptage du neurone i, d'intensité $\lambda_t^{(i)}$ qui dépend de tous les temps de sauts antérieurs à t). Cette modélisation permet de prendre en compte la dépendance temporelle forte existante dans un train de spikes et aussi entre les trains de spikes de plusieurs neurones. Nous utilisons deux méthodes d'estimation de l'intensité des processus de Hawkes linéaires, l'une paramétrique appelée ADM4 (Zhou *et al.*(2013)), basée sur la vraisemblance d'un Hawkes linéaire multidimensionnel de noyau exponentiel et adaptée à la grande dimension des paramètres grâce à une pénalité de type LASSO. Elle est implémentée dans la documentation Python `tick` que nous utilisons. La seconde méthode non-paramétrique de Lambert *et al.*(2018) est basée sur un critère de moindres carrés avec une pénalité de type LASSO également, implémentée dans R. L'idée de cet estimateur peut être rapprochée de celle du cross-correlogram ou du PSTH dans le sens où les fonctions noyaux du processus de Hawkes sont constantes par morceaux.

Ces deux méthodes présentent des avantages et des inconvénients. En effet, la méthode ADM4 est adaptée à la grande dimension mais repose d'une part sur l'hypothèse d'un noyau exponentiel et également elle ne permet de détecter que des phénomènes d'excitation. La seconde méthode, non-paramétrique, est plus robuste à la forme des interactions et également elle permet de détecter des phénomènes d'inhibition. Cependant, cette méthode est mal adaptée à la grande dimension comme précisé dans l'article Lambert *et al.*(2018).

On applique les deux méthodes aux trains de spikes de 25 neurones (sélectionnés à l'aide du PSTH), le 26ème neurone étant le neurone central (dont on a extrait le train de spike à partir du signal continu intracellulaire). Les sorties des algorithmes sont très proches : on voit notamment des coefficient non nuls sur presque toutes les entrées de la diagonale, ce qui correspond à un phénomène d'auto-excitation pour chacun de ces neurones. On détecte également des interactions entre les différents neurones mais la matrice d'adjacence reste parcimonieuse, ce qui permet notamment d'extraire un sous-

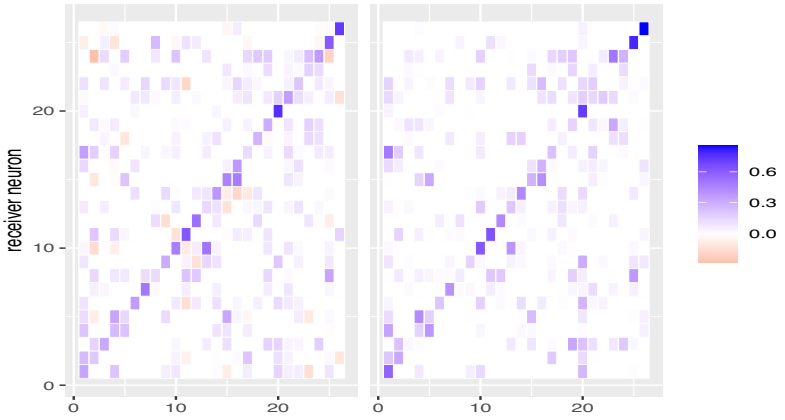


Figure 1: Représentation des matrice d'adjacence sur 26 neurones. L'entrée (i, j) qui correspond à l'effet du neurone i sur le neurone j. Gauche: méthode non-paramétrique de Lambert *et al.* (2018). Droite: Méthode ADM4.

ensemble de neurones qui interagissent avec le neurone central. Enfin, on remarque que la méthode non-paramétrique permet de détecter des phénomènes d'inhibition qui, comme attendu, ne sont pas visibles avec la méthode ADM4.

4 Validation du réseau

Le but de cette partie est double. Nous souhaitons vérifier avec les mesures intracellulaires, que les temps de spikes des neurones qui impactent le neurone central selon nos estimations, impactent la trajectoire du potentiel de membrane de ce neurone. Puis nous voulons valider l'hypothèse "Hawkes" sur ces données.

Validation à l'aide du signal intracellulaire D'après l'estimation de la matrice d'adjacence, on lit que les neurones principaux agissant sur le neurone 26 (le neurone central) sont les neurones 4 et 20 d'après ADM4. Sur la Figure 2, on a représenté, pour le premier échantillon, la trace du potentiel de membrane sur le graphe du haut, et en bas, les trains de spikes des deux neurones impactant le neurone central (avec entre parenthèse l'estimation du coefficient d'interaction du neurone sur le neurone central). On voit clairement le lien entre les temps de spikes et les variations du signal continu.

Validation du modèle Pour la validation du modèle de Hawkes nous avons utilisé deux méthodes. La première est la méthode classique du "Time rescaling theorem" de Brown (2002) utilisée par exemple par Pouzat & Chaffiol (2009). Cette technique utilise le fait que pour un processus ponctuel stationnaire, le compensateur $\Lambda_t = \int_0^t \lambda_s ds$ évalué

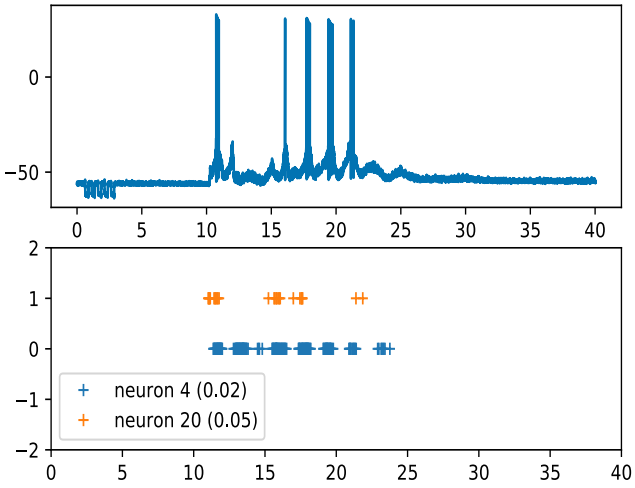


Figure 2: (Échantillon 1) Haut: trace du potentiel de membrane. Bas: train de spikes pour les neurones 4 et 20.

au temps de sauts du processus, suit un processus de Poisson d'intensité 1. Dans notre cas nous regardons ce processus pour chaque neurone. Cependant dans l'article Reynaud-Bouret *et al.*(2014) il a été soulevé que le plug-in d'un estimateur de l'intensité dans cette procédure induit de mauvais résultats dans le test de Kolmogorov-Smirnov. Il est alors conseillé de sous-échantillonner. Nous appliquons donc le test 4 de l'article. Cependant, les garanties obtenues pour ce test sont asymptotiques en la taille de l'échantillon et dans notre expérience nous disposons de 10 échantillons seulement.

5 Modélisation d'un intervalle inter-spikes du potentiel de membrane

L'étude préliminaire présentée ci-dessus va nous permettre de modéliser le potentiel de membrane du neurone central. En effet, une fois extrait un petit réseau de neurones communiquant avec le neurone central, nous proposons de modéliser le potentiel de membrane de ce neurone entre deux temps de spikes. Nous supposons pour cela que la dynamique du potentiel pendant ces phases est décrite par l'équation suivante

$$dX_t = b(X_t) + \sigma(X_t)dW_t + a(X_{t-}) \sum_{j=1}^{M} dN_t^{(j)}$$

où W est un mouvement Brownien standard et N un processus de Hawkes multivarié indépendant de W. Les travaux de Dion *et al.*(2019) et Amorino *et al.*(2020) permettent d'estimer les paramètres b, σ, a. Enfin, en simulant un grand nombre de trajectoires à partir des estimations, nous obtenons un intervalle de confiance empirique et vérifions que les trajectoires (non utilisées pour l'estimation) sont bien dans cet intervalle.

Bibliographie

Amorino, C., Dion, C., Gloter, A. and Lemler, S. (2020). On the nonparametric inference of coefficients of self-exciting jump-diffusion. *arXiv preprint arXiv:2011.12387*

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear hawkes processes. *The Annals of Probability* pp. 1563-1588.

Brown, E., Barbieri, R., Ventura, V., Kass, R. and Frank, L. (2002) The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*

Daley, D. and Vere-Jones, D. (2007). An introduction to the theory of point processes: volume II: general theory and structure. *Springer Science & Business Media.*

Dion, C. and Lemler, S. (2019). Nonparametric drift estimation for diffusions with jumps driven by a Hawkes process. *Statistical Inference for Stochastic Processes.*

Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., and Reynaud-Bouret, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of neuroscience methods*

Radosevic, M., Willumsen, A., Petersen, P., Lindén, H., Vestergaard, M. and Berg, R. (2019) Decoupling of timescales reveals sparse convergent CPG network in the adult spinal cord. *Nature communications*

Reynaud-Bouret, P., Rivoirard, V., Grammont, F. and Tuleau-Malot, C. (2014) Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*

Pouzat, C., & Chaffiol, A. (2009). On goodness of fit tests for models of neuronal spike trains considered as counting processes. *arXiv preprint arXiv:0909.2785*

Zhou, K., Zha, H. and Song, L. (2013) Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. *Artificial Intelligence and Statistics*

Algorithme d'ensembles actifs par fenetre glissante pour l'estimation parcimonieuse de modèle convolutionel

Laurent Dragoni [1] & Karim Lounici [2] & Rémi Flamary [2] & Patricia Reynaud-Bouret [1]

[1] *Université Côte d'Azur, CNRS, Laboratoire J.A. Dieudonné, 06108 Nice, France*
[2] *Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau*

Résumé. Nous présentons un algorithme de résolution rapide du Lasso dans le cadre de modèles convolutionnels en grande dimension. Des simulations numériques illustrent l'efficacité de notre approche. De plus, nous démontrons théoriquement que la complexité temporelle de cet algorithme croît linéairement avec la taille du signal enregistré.

Mots-clés. Parcimonie, Lasso, Optimisation, Neurosciences, Tri de potentiels d'action

Abstract. We present a fast algorithm for the resolution of the Lasso for convolutional models in high dimension. Numerical simulations illustrate the efficiency of our approach. Moreover we show theoretically that the temporal complexity of this algorithm growths linearly w.r.t the size of the recorded signal.

Keywords. Sparsity, Lasso, Optimization, Neurosciences, Spike sorting

1 Introduction

Nous proposons un nouvel algorithme de résolution du Lasso pour l'étude de modèles convolutionnels. Sous une hypothèse de parcimonie du vecteur à estimer, nous affinons la stratégie dite d'ensembles actifs ou *active set* en un algorithme en ligne performant en grande dimension. Nous montrons de plus que la complexité temporelle théorique de cet algorithme croît linéairement avec la taille du signal enregistré. Cet algorithme générique peut s'appliquer à divers domaines, comme notamment au problème du tri de potentiels d'action en neurosciences dont la problématique porte sur l'estimation des formes des potentiels d'action et des instants d'activation des différents neurones à partir de l'enregistrement de l'activité neuronale.

Modèle convolutionnel Durant une expérience, d électrodes enregistrent l'activité de q neurones. Chaque électrode enregistre un signal de taille n (nombre de pas de temps). Nous proposons de de mettre en relation l'activité des neurones et les signaux enregistrés en utilisant un modèle convolutionnel, déjà proposé par Roberts (1979). Ce modèle s'écrit sous la forme

$$\boldsymbol{S} = \sum_{j=1}^{q} \boldsymbol{W}_j * \boldsymbol{a}_j + \boldsymbol{N}, \tag{1}$$

où $\boldsymbol{S} \in \mathbb{R}^{d\times n}$ est la matrice des observations, contenant les d signaux enregistrés de taille n. La matrice $\boldsymbol{W}_j \in \mathbb{R}^{d\times \ell}$ contient les formes des potentiels d'action du neurone j sur toutes les électrodes. Notons que toute forme de potentiel d'action est décrite par ℓ points. Le vecteur $\boldsymbol{a}_j \in \mathbb{R}^n$ est appelé le vecteur d'activation du neurone j. Les entrées non nulles de $\boldsymbol{a}_j$ correspondent aux instants d'activation de ce neurone. Étant donné que la fréquence de décharge des neurones est très petite devant la fréquence d'acquisition du signal, notons que $\boldsymbol{a}_j$ est un vecteur sparse. L'opérateur de convolution par rapport au temps est noté $*$ (d'où le nom de modèle *convolutionnel*). Finalement, $\boldsymbol{N} \in \mathbb{R}^{d\times n}$ est une matrice de bruit.

La convolution étant un opérateur linéaire, le modèle convolutionnel (1) se reformule en un modèle linéaire, après une étape de vectorisation. On obtient alors le problème suivant

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{\sigma}, \tag{2}$$

où $\boldsymbol{H}$ est une matrice bloc-Toeplitz de taille $dn \times qn$ codant la convolution entre les formes des potentiels d'action et le vecteur d'activation $\mathbf{a}$. Les potentiels d'action étant de longueur ℓ très petite devant n, la matrice $\boldsymbol{H}$ est en pratique extrêmement sparse.

Dans la suite, on s'intéresse à l'estimation du vecteur d'activation $\mathbf{a}$ en supposant connus le nombre de neurones q et les formes des potentiels d'action $\boldsymbol{W}_j$. Cet objectif correspond à la deuxième étape de la stratégie de tri de potentiels d'action plus générale suivante:

1. estimation du nombre de neurones et des formes des potentiels d'action
2. identification et classification des événements présents dans le signal

Lasso et conditions d'optimalité Estimer $\boldsymbol{a}$ quand le nombre de neurones est supérieur au nombre d'électrodes serait impossible sans hypothèse structurelle supplémentaire. Comme annoncé en section 1, une propriété importe du problème est que les neurones ont tendance à peu décharger. Par conséquent, le nombre de coefficients non nuls dans $\boldsymbol{a}$ est petit devant qn. Ceci nous invite à considérer un estimateur promouvant la parcimonie, comme le Lasso proposé par Tibshirani (1996).

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}\in\mathbb{R}^{qn}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|_2^2 + \lambda \|\boldsymbol{a}\|_1 , \tag{3}$$

où $\lambda > 0$ est l'unique paramètre de la méthode et qui dépend du ratio signal sur bruit. Une propriété cruciale du Lasso est la condition d'optimalité suivante : soit $\hat{a}$ une solution de (3) et écrivons $\boldsymbol{H}_j$ la j-ème colonne de $\boldsymbol{H}$, où $1 \leq j \leq qn$.

$$\forall j, \quad \text{si} \quad |\boldsymbol{H}_j^T(\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{a}})| < \lambda, \quad \text{alors} \quad \hat{\boldsymbol{a}}_j = 0. \tag{4}$$

Algorithme d'ensembles actifs (active set) Le calcul d'un estimateur Lasso, c'est-à-dire d'une solution du problème (3), peut être très coûteux en grande dimension. Nous exploitons alors la stratégie dite de l'*active set* nous permettant de calculer de manière plus efficace le Lasso. Notons que cette stratégie assez générique a déjà été exploitée dans d'autres contextes, cf. Lee et al. (2007), Szafranski et al. (2008) et Boisbunon et al. (2014). Ici l'idée principale de l'*active set* repose sur l'exploitation des conditions d'optimalité (4). Initialisant l'estimateur Lasso au vecteur nul, l'objectif est d'activer de manière itérative ses coordonnées j ne vérifiant pas (4), tout en mettant à jour l'estimateur Lasso à chaque activation de coordonnées. On appelle active set et on note J l'ensemble de ces coordonnées actives. Le vecteur $\boldsymbol{a}$ étant très sparse, on s'attend ainsi à résoudre des problèmes Lasso de taille $|J|$ très petite devant qn. Dans le pire des cas, l'algorithme de l'active set active toute les coordonnées possibles, ce qui conduit à calculer une solution Lasso sur l'espace tout entier. Ainsi, l'algorithme de l'active set termine toujours en temps fini.

2 Active set par fenêtre glissante adaptative

En pratique, le calcul des conditions d'optimalité et la mise à jour du Lasso sur J sont les étapes les plus coûteuses de l'active set. Au vu des dimensions du problème, ceci ne permet pas encore de calculer rapidement une solution en temps raisonnable. Nous proposons alors l'idée de l'active set par fenêtre glissante, qui exploitera davantage la structure du problème. Une notion cruciale est celle de recouvrement temporel entre activations : les activations étant rares et leurs effets sur le signal étant très localisés, deux activations suffisament distantes n'ont aucune influence réciproque. Plus précisément, rassemblons les temps d'activation de tous les neurones dans un vecteur noté $\boldsymbol{a}^{times}$ et considérons alors $\boldsymbol{a}_i^{times}$ et $\boldsymbol{a}_j^{times}$ deux activations successives de a^{times}. Si on a $|\boldsymbol{a}_i^{times} - \boldsymbol{a}_j^{times}| \leq \ell$, on dira que ces deux activations se recouvrent. On appelle alors recouvrement de l'activation $\boldsymbol{a}_i^{times}$ sa composante connexe pour la relation qui précède. On partionne ainsi l'ensemble des temps d'activations en des recouvrements disjoints. L'idée essentielle de l'active set par fenêtre glissante est de retrouver ces recouvrements en parcourant le domaine temporel à l'aide de fenêtres dont la taille peut évoluer et donc d'exploiter cette séparation des activations afin de résoudre des problèmes indépendants et de taille plus petite.

Retrouver les recouvrements des activations exige une évolution précise des fenêtres de l'algorithme. Une fois les conditions d'optimalité satisfaites sur ω (ligne 3 de algorithm 1), si le signal reconstruit est entièrement contenu dans ω, autrement dit si tous les temps d'activation sont à distance supérieure à ℓ du bord j, alors le recouvrement est entièrement contenu dans ω. Or on sait que la résolution du Lasso pour le recouvrement suivant est indépendante du recouvrement actuel. Dans ce cas, on a donc trouvé la solution Lasso pour la fenetre actuelle peut donc travailler sur une nouvelle fenetre immédiatement après. Sinon, on étend la fenêtre actuelle et on résout à nouveau le Lasso sur celle-ci en mettant

Algorithm 1 Structure de l'active set par fenêtre glissante

Initialisation de la fenetre $\omega = [\![i, j]\!] = [\![1, \eta]\!]$
repeat
  $\boldsymbol{a}_\omega$ = solution du Lasso sur ω via l'active set
  if Le signal reconstruit est contenu dans ω **then**
    Stockage de la solution $\boldsymbol{a}_\omega$, décalage de la fenêtre $\omega = [\![j+1, j+\eta]\!]$
  else
    Extension de la fenêtre actuelle $\omega = [\![i, j+\ell]\!]$
  end if
until $j = n$ // Fin du signal

à jours de manière itérative avec l'*active set*. Le vecteur précédemment calculé complété par des zéros constitue alors une initialisation raisonnable pour la résolution du nouveau Lasso.

Dans l'*active set* standard, le coût du calcul des conditions d'optimalité est de l'ordre de $O(nqd\ell)$ à chaque étape. Dans l'*active set* par fenêtre glissante, ce coût se réduit donc à $O(|\omega|qd\ell)$, avec $|\omega| << n$.

3 Expérimentations numériques

Afin d'illustrer les performances de l'algorithme de l'active set par fenêtre glissante, nous présentons ici une comparaison des temps d'exécution pour trois approches différentes : l'approche frontale qui consiste à résoudre le Lasso (3) globalement, l'active set générique et l'active set par fenêtre glissante. Quelle que soit l'approche, nous résolvons les Lasso en utilisant l'algorithme du gradient proximal accéléré, implémenté dans FISTA par Beck et Teboulle (2009). Nous avons simulé notre jeu de données de manière réaliste en utilisant notamment le modèle classique de Hodgkin et Huxley (1952) pour la description des formes de potentiels d'action et implémenté par Pouzat (2016). Afin de favoriser l'étude l'influence de n, nous nous sommes limité à des valeurs raisonnables pour le nombre de neurones ($q = 5$) et d'électrodes ($d = 4$).

FISTA global et l'active set générique nécessitent une grande utilisation de la mémoire pour le stockage de la matrice $\boldsymbol{H}$ dont la taille augmente en $O(n^2)$. Ainsi les simulations pour ces deux méthodes deviennent rapidement prohibitives. Nous constatons néanmoins en figure 1 que l'active set par fenêtre glissante est visiblement plus rapide que ces deux méthodes. En effet, celui-ci semble croître linéairement en n. De plus, il nécessite une occupation de la mémoire nettement plus raisonnable.

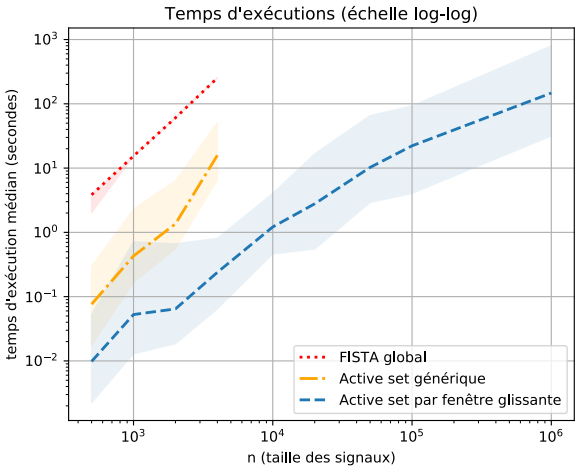


Figure 1: Temps d'exécution des algorithmes pour différentes valeurs de n.

4 Étude mathématique de l'algorithme et complexité

Étant donné la structure de l'algorithme par fenêtre glissante, une question naturelle apparaît : peut-on s'assurer que la solution $\hat{\boldsymbol{a}}$ calculée est bien une solution du problème initial ? Par construction, $\hat{\boldsymbol{a}}$ est obtenu par recollements successifs de solutions sur des fenêtres disjointes. Nous pouvons répondre par l'affirmative.

Theorem 1. *La solution $\hat{\boldsymbol{a}}$ calculée par l'active set par fenêtre glissante est une solution Lasso du problème initial, c'est-à-dire de* (3).

Dans la suite, on s'intéresse à l'estimation de la complexité algorithmique de l'active set par fenêtre glissante. On appelle fenêtre *maximale* toute fenêtre qui a été quittée à l'issue de l'étape (5) de l'algorithme 1.

Lemma 1. *On note respectivement I^* et $\hat{I}$ l'ensemble des coordonnées non nulles des vecteurs a et $\hat{a}$. Alors sur l'événement $I^* = \hat{I}$, toute fenêtre maximale contient au plus un recouvrement. De plus, si une fenêtre maximale contient un recouvrement de taille k, alors cette fenêtre est de longueur au plus $k + \eta + \ell$.*

Utilisant des résultats théoriques sur le Lasso, comme Bunea (2008), nous pouvons montrer que la condition $I^* = \hat{I}$ est satisfaite avec grande probabilité. Le lemme précédent nous informe donc bien sur le lien entre la complexité de l'algorithme (la taille des fenêtres) et la taille des recouvrements. Pour des raisons techniques, nous étendons ici la distance de détection des recouvrements à 3ℓ.

Lemma 2. *Supposons que le vecteur des temps d'activation $\boldsymbol{a}^{times}$ soit tiré selon un processus de Bernoulli de probabilité p. Alors la taille moyenne d'un recouvrement est inférieure à $3\ell(1-p)^{-3\ell}$.*

Les résultats obtenues en section 3 semblaient indiquer que la complexité temporelle de l'active set par fenêtre glissante croît en $O(n)$. Nous sommes ici en mesure de le démontrer mathématiquement.

Theorem 2. *La complexité temporelle moyenne de l'algorithme de l'active set par fenêtre glissante est de l'ordre*

$$O(n\bar{\omega}^4 q^2 (d+q)^2), \tag{5}$$

où $\bar{\omega}$ est la taille moyenne d'une fenêtre maximale.

Les lemmes précédents nous assurent alors que $\bar{\omega}$ est indépendant de n. Ceci est à notre connaissance le premier résultat théorique sur la **résolution d'un Lasso structuré avec une complexité** $O(n)$. Notons que le résultat du théorème 2 est pessimiste. Dans des travaux futurs, nous nous attacherons à prendre en compte la dimension spatiale du problème, ce qui permettrait de résoudre des sous-problèmes associés à de petits sous-ensembles de neurones. Auquel cas les termes d et q apparaissant dans (5) seraient remplacés par des valeurs bien inférieures.

Bibliographie

Beck, A., & Teboulle, M. (2009). *A fast iterative shrinkage-thresholding algorithm for linear inverse problems.* SIAM journal on imaging sciences, 2(1), 183-202.

Boisbunon, A., Flamary, R., Rakotomamonjy, A., Giros, A., & Zerubia, J. (2014, September). *Large Scale Sparse Optimization for Object Detection in High Resolution Images.*

Bunea, F. (2008). *Consistent selection via the Lasso for high dimensional approximating regression models.* In Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh (pp. 122-137). Institute of Mathematical Statistics.

Hodgkin, A. L., & Huxley, A. F. (1952). *A quantitative description of membrane current and its application to conduction and excitation in nerve.* The Journal of physiology, 117(4), 500-544.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). *Efficient sparse coding algorithms.* In Advances in neural information processing systems (pp. 801-808).

Pouzat, C. (2016). *Origin of the high frequency extra-cellular signal.* http://christophe-pouzat.github.io/LASCON2016/OriginOfTheHighFrequencyExtraCellularSignal.html.

Szafranski, M., Grandvalet, Y., & Morizet-Mahoudeaux, P. (2008). *Hierarchical penalization.* In Advances in neural information processing systems (pp. 1457-1464).

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Roberts, W.M. (1979) *Optimal recognition of neuronal waveforms.* Biol. Cybernetics 35, 73–80 .

Unsupervised classification of spectra of galaxies: the ICL criterion, strength and weakness

Julien Dubois [1] & Didier Fraix-Burnet [1] & Charles Bouveyron [2] & Jihane Moultaka [3] & Pooja Sharma [4] & Denis Burgarella [4]

[1] *Univ. Grenoble Alpes, CNRS, IPAG, Grenoble, France*
(julien.dubois@univ-grenoble-alpes.fr, didier.fraix-burnet@univ-grenoble-alpes.fr)
[2] *Université Côte d'Azur, Inria, CNRS, LJAD, Maasai, Nice, France*
(charles.bouveyron@univ-cotedazur.fr)
[3] *IRAP, Université de Toulouse, CNRS, CNES, UPS, 14, avenue Edouard Belin, F-31400 Toulouse, France*
[4] *Aix-Marseille Université, CNRS, LAM (Laboratoire d'Astrophysique de Marseille) UMR 7326, 13388 Marseille, France*

Résumé. Le traitement d'une grande quantité de données est une nouvelle tâche problématique en astrophysique. La réduction de la dimension du nombre d'observations et du nombre de variables (caractéristiques, observables) est nécessaire pour une compréhension physique plus facile. Dans cet exposé, nous présentons plusieurs résultats obtenus avec Fisher-EM, un algorithme de classification non supervisé en modèle de mélange dans un sous-espace latent discriminant, appliqué sur plusieurs échantillons de spectres de galaxies : deux échantillons simulés et un grand échantillon observé (702248 spectres de 1437 longueurs d'onde chacun). Nous sommes ainsi en mesure de démontrer la faisabilité, la robustesse et l'utilité de la classification non supervisée pour de très grands échantillons de spectres de galaxies.

Traduit avec www.DeepL.com/Translator (version gratuite) **Mots-clés.** Classification non supervisée, Machine Learning, Spectres, Astrophysique, Galaxies

Abstract. Dealing with large amount of data is a new problematic task in astrophysics. Dimensionality reduction in both the number of observations and the number of variables (features, observables) is necessary for an easier physical understanding. In this talk, we present several results obtained with Fisher-EM, an unsupervised clustering discriminative latent mixture model algorithm, applied on several samples of spectra of galaxies: two simulated samples and a large observed one (702248 spectra of 1437 wavelengths each). We are thus able to demonstrate the feasibility, robustness and usefulness of unsupervised classification of very large samples of spectra of galaxies.

Keywords. Unsupervised classification, Machine Learning, spectra, astrophysics, galaxies

1 Introduction

Astrophysics has now entered the era of Big Data and the new telescopes and instruments that will come into operation in the next few years (EUCLID, VLT/MOONS, LSST, SKA...) face technological challenges for the management and the analysis of the data. Spectra are particularly spectacular since they contain several thousands of wavelengths making matrices of about a million observations described by thousands of parameters.

These spectra contain all the astrophysical information that an astronomer can dream of, apart from the morphological structure: the composition of the stellar populations, the history of the stellar formation events, the content in gas and its physical conditions, the presence of hot regions such star forming regions, hot nebulae, active galactic nuclei hosting black holes, and the global kinematics of the galaxy. Basically the spectrum of a galaxy is made of a continuum due to the thermal emission from the stars, plus some absorption features due to the cold gas, and emission lines due to hot gas. An atlas of typical galaxy spectra is provided in Kennicutt (1992) or Dobos et al. (2012).

Classification in astrophysics traditionally uses an eye-based approach which also serves as the basis for supervised learning studies. In contrast, unsupervised learning is not common (Fraix-Burnet et al. 2015). In the case of spectra of galaxies, the first study using k-means is recent (SÀnchez Almeida et al. 2010) and has been disputed in De et al. (2016). Indeed, the data set is so large that this simple technique is not able to detect structures.

In the present study, we instead use a more sophisticated algorithm and demonstrate its capacity to discriminate physically distinct galaxies from their optical spectra even with very large samples. Two other samples of simulated spectra of galaxies allow us to understand how the algorithm uses the spectra to build distinct classes as well as the effect of the noise on the final classification.

2 Fisher-EM in short

The unsupervised clustering was performed with the algorithm Fisher-EM available in R (Bouveyron and Brunet 2012). The Fisher-EM algorithm is a discriminative latent mixture model that estimates both the discriminative subspace and the parameters of the mixture model. It is based on the Expectation-Maximization (EM) algorithm from which an additional step, named F-step, is introduced, between the E- and the M-step. This F-step uses the Fisher criterion under orthonormality constraints and conditionally to the posterior probabilities to optimize the clustering.

3 A sample of Single Stellar Population spectra

The most simple spectrum comes from a single stellar population (set of stars of same age and same initial chemical composition) without considering the emission lines from the ionized gas surrounding them. 966 spectra with 3563 wavelengths each were simulated. Fisher-EM finds an optimum of 105 clusters with very little dispersion within each.

Only three parameters were used to simulate these spectra: age, metallicity and the initial mass function. The 105 classes are very homogeneous and discriminate very well both in the spectra space and the parameter space. Remarkably, the Fisher-EM algorithm is able to mainly resolve the degeneracy between age and metallicity that both affect a spectrum in a similar way.

3.1 The CIGALE sample: a simulated set of more complete spectra

The CIGALE code (Boquien et al 2019) can simulate many physical phenomena that constitutes the spectrum of a galaxy. Such a code is used to fit observed spectra or to create mock catalogues. We have simulated 11000 different spectra in the optical range with 494 wavelengths each, by varying 12 parameters (star formation history, dust...). This sample was used to investigate the convergence frequency to find the best number of clusters according to the ICL criterion, as well as the effect on the noise on the resulting classification.

A very difficult problem in unsupervised clustering is the search for the best solution. With very large samples, computation time is a severe limitation. The Fisher-EM analysis of the CIGALE sample has shown that repeating the analysis many times is not required when the solution is already good. This means that the curve of the ICL versus the number of clusters may not necessarily point to the best solution (if it exists), but provides a very reasonable classification.

The effect of the noise was found to be somewhat dependent on the details within the spectra rather than on the signal-to-noise ratio. Adding even a very small noise greatly affects the classification scheme for the CIGALE spectra, contrarily to the SSP spectra which have no emission lines.

3.2 The SDSS sample: a very large observed data set

A very large sample of 702248 spectra with 1437 wavelengths in the optical range was retrieved from the Sloan Digital Sky Survey (SDSS) database. Since many runs are required to find the best statistical model and the optimum number of clusters, we adopted a strategy that first explored the data with subsets of 100000 spectra, and then considered a 302248 subset for the final classification, with a distinct 300000 sample used as a validation.

The first exploration imposed a clustering in several steps: one or two preliminary steps with K=3 and then the sub-clustering of the three classes. The reason is that increasing the number of clusters for the preliminary steps did not change much the distribution of the three larger clusters, but increases the number of very small clusters (with down to one element).

This sub-clustering strategy revealed itself quite robust among the different subsets. Because of excessive computation time, only the preliminary steps were performed on the full sample.

However, this strategy shows that the analyses of several subsets of the full sample helps find a representative sub-sample so that the resulting classification proved to be robust.

The optimum number of clusters is found to be 86 and are shown in Fig. 1. The homogeneity of most of the classes is striking. The number of very small classes is large, 37 classes gathering 99% of the sample. Some small classes are made of clearly suspicious spectra, showing the capacity of Fisher-EM to distinguish outliers and artifacts.

A first interpretation of the classes demonstrates the astrophysical relevance of these classes.

4 Conclusion

The thorough study of the unsupervised classification of large samples of spectra of galaxies have demonstrated that this approach is suited to cope with the avalanche of data that is coming in astrophysics.

We have developed a strategy to analyse any sample that would prove otherwise to be computationally impossible to tackle. More importantly, the resulting classification is shown to be robust, discriminative and relevant.

We intend to fully understand the physics of the galaxies within each class. We will try to generalize this classification to the millions of other spectra available in present and future databases, in order to obtain a general atlas of the diversity of galaxies in the Universe.

Bibliographie

Boquien, M., Burgarella, D., Roehlly, Y., et al. (2019). CIGALE: a python Code Investigating GALaxy Emission. *Astronomy & Astrophysics*, 622, A103

Bouveyron, C. and Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22, 301

Fraix-Burnet, D, Bouveyron, C. and Moultaka, J. (2021). Unsupervised classification of SDSS galaxy spectra. *Astronomy & Astrophysics*, in press.

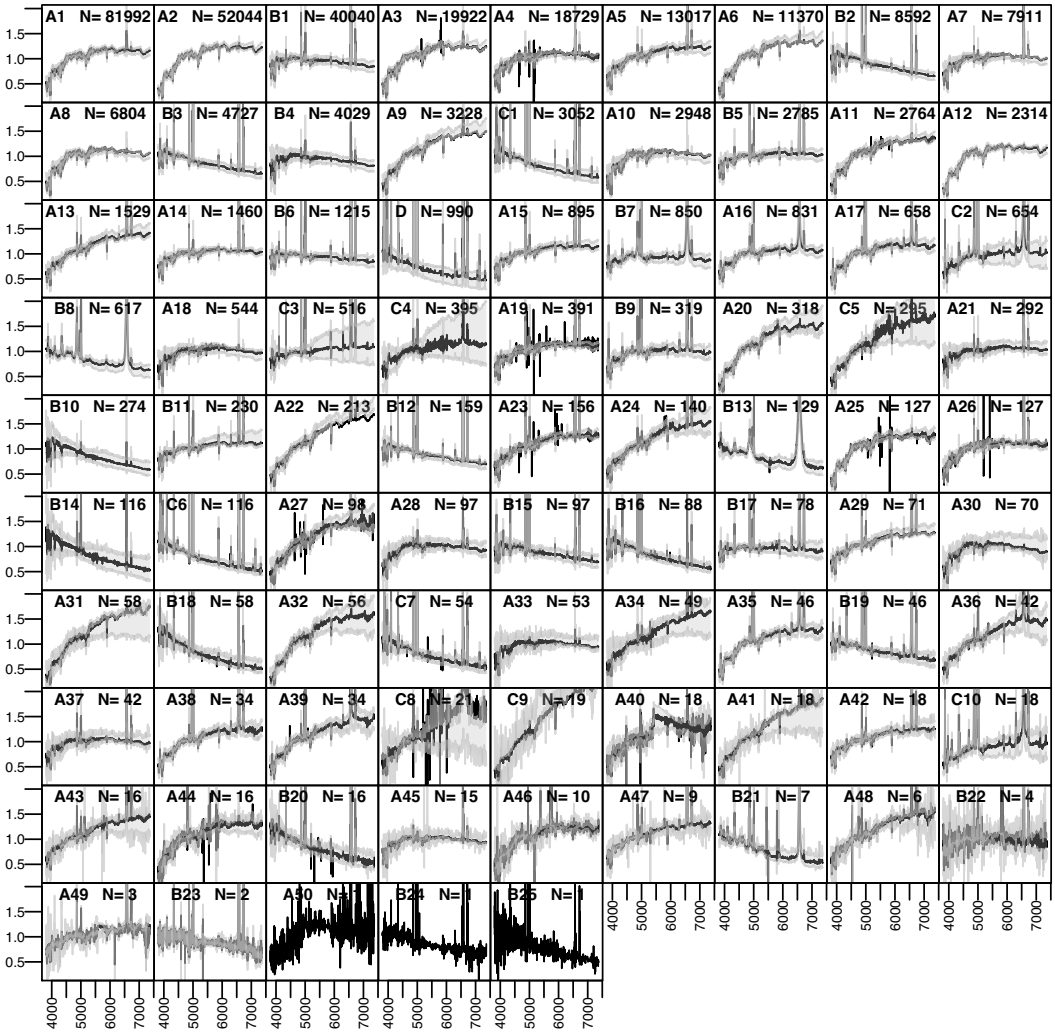


Figure 1: The resulting 86 classes for the SDSS sample stored in decreasing order of the number of spectra within a class. The black line is the mean spectrum of each class, and the grey zone lies between the 10 and 90% quantiles for that class. The vertical scale is arbitrary and identical for all spectra. The class index and the number N of objects in the class are given in each graph (from Fraix-Burnet et al 2021).

Une approche de modélisation par équations structurelles pour l'étude de la causalité en agroécologie

Mathieu Emily [1] & Sébastian Mira [2] & Edith LeCadre [3]

[1] *Institut Agro, CNRS, Univ Rennes, IRMAR (Institut de Recherche Mathématique de Rennes) - UMR 6625, F-35000 Rennes, France*

[2] *SAS, INRAE, Institut Agro, 65 rue de Saint-Brieuc CS 84215, 35042, Rennes Cedex, France – EUREDEN Innovation, ZI Lanrinou, CS 20100, 29 206 Landerneau Cedex, France*

[3] *SAS, INRAE, Institut Agro, 65 rue de Saint-Brieuc CS 84215, 35042, Rennes Cedex, France*

Résumé. La modélisation par équations structurelles permet l'estimation et l'identification de relations causales entre des composants d'un système d'étude. Des approches parcimonieuses permettent de modifier un schéma causal en enlevant certaines relations tout en conservant la totalité des variables du système. Toutefois, dans certains domaines comme l'agroécologie par exemple, la multiplicité des mesures peut devenir très couteuse et il devient intéressant de pouvoir sélectionner les variables suffisantes pour l'identification de relations causales. Pour répondre à cette problématique, nous proposons une statistique de test définie comme la différence de mesure d'ajustement entre deux modèles concurrents. En utilisant un principe de permutations, nous avons pu tester l'apport des données de biomasses microbiennes et montrer que celles-ci ne jouaient pas un rôle significatif dans le système causal.

Mots-clés. Modèles à équations structurelles, Sélection de variables, Agroécologie

Abstract. Structural equations modeling allows the estimation and identification of causal relationships between components of a system. The use of parsimonious approaches make it possible to modify a causal pattern by removing certain relations while retaining all the variables of the system. However, in fields such as agroecology for example, the multiplicity of measures can become very expensive and selecting subset of variables, sufficient for the causal system, is of particular interest. To tackle this issue, we propose a test statistic defined as the difference in fit measurement between two competing models. By using permutations, we were able to test the contribution of microbial biomass data and show that they did not play a significant role in the causal system.

Keywords. Structural equation models, Variable selection, Agroecology

1 Introduction

L'étude des relations causales entre les éléments d'un système complexe, par le biais de l'inférence causale, a donné lieu à une abondante littérature ces dernières années. Récemment, la publication d'un numéro spécial dans le journal de la société française de statistique présente les différentes définitions de la causalité ainsi que les méthodes pour identifier et estimer les relations causales (Benkeser et al., 2020). Parmi les familles de modèles statistiques utilisées pour l'inférence causale, nous nous intéressons ici aux modèles à equations structurelles (SEM) qui permettent la modélisation du système d'étude dans sa globalité. Le principal avantage des SEM réside dans la formalisation de l'ensemble des relations causales ainsi que dans la visualisation du système causal par un graphe orienté acyclique (DAG). Dans leur feuille de route pour l'étude de la causalité, Petersen et van der Laan proposent de définir et de fixer le modèle causal avant toute confontration aux données (Petersen and van der Laan, 2014). Dans la suite de la feuille de route, la démarche statistique s'intéresse à modifier certaines relations causales soit en rajoutant certains liens, notamment à l'aide des indices de modifications, soit en favorisant la parcimonie du modèle par des méthodes de régularisation (Jacobucci et al. 2016) ou d'une modélisation sparse (Cai et al. 2013).

Dans toutes ces études, l'amélioration du modèle structural se fait sans modification de l'ensemble des variables mesurées. Pourtant, dans de nombreux domaines, l'obtention de certaines variables peut être très couteuse. Dans cet article, nous présentons une méthode permettant de déterminer si une variable observée a un apport significatif dans la qualité de la modélisation du système causal. Cette méthode s'appuie sur la mesure d'ajustement de la vraisemblance dont la significativité est testée par une permutation des données.

Ce travail est motivé par l'amélioration des pratiques agricoles dans un contexte agroécologique. Afin d'étudier le rôle de la composition du sol dans les performances aériennes du blé, un ensemble de 6 mesures ont été effectuées en conditions controlées sur 55 plans de blé. Parmi ces 6 mesures, nous retrouvons 3 mesures d'activités enzymatiques (PHO pour l'acide phosphatase, PAK pour l'alkaline phosphatase et GLU pour la β-glucolidase), une mesure de biomasse microbienne (Bmc), la biomasse racinaire (BSr) et le nombre de talles (Ntl). Dans une première étape, un modèle structural est confronté aux données récoltées et dans un second temps, l'importance de la mesure de biomasse microbienne est testée par notre méthode.

2 Modèle à équations structurelles

2.1 Notations

D'après Bollen (1989), un modèle général d'équations structurelles s'appuie sur un ensemble de $p+q$ variables observées (ou variables manifestes). Les variables observées peuvent se décomposer en deux sous-ensembles : un sous-ensemble $\mathbf{x}$ de p variables exogènes

(*i.e.* définies en dehors du système) et un sous-ensemble $\mathbf{y}$ de q variables endogènes (*i.e.* définies par le système étudié). En outre, un ensemble de $m+n$ variables latentes (m variables latentes endogènes et n variables latentes exogènes) sont également définies à travers le schéma conceptuel du système. De nouveau, l'ensemble des variables latentes peut se décomposer en un ensemble η de m variables endogènes et un ensemble ξ de n variables exogènes. Ainsi le modèle général est défini par deux sous modèles : le modèle structurel (voir équation (1)) et un modèle de mesures (voir équations (2) et (3)).

Le modèle structurel peut donc s'écrire de la façon suivante :

$$\eta = \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta \tag{1}$$

où $\mathbf{B}$ (de taille $m \times m$) et $\mathbf{\Gamma}$ (de taille $m \times n$) correspondent aux matrices des coefficients modélisant le lien entre les variables latentes. ζ un vecteur d'erreur (avec $\mathbb{E}[\zeta] = 0$ et $cov(\zeta, \xi) = 0$)). Le modèle de mesure s'écrit de la façon suivante :

$$\mathbf{y} = \mathbf{\Lambda_y}\eta + \varepsilon \tag{2}$$

$$\mathbf{x} = \mathbf{\Lambda_x}\xi + \delta \tag{3}$$

où $\mathbf{\Lambda_y}$ (de taille $p \times m$), $\mathbf{\Lambda_x}$ (de taille $q \times n$) sont les matrices des coefficients modélisant le lien entre les variables observées et variables latentes. Les vecteurs ε et δ représentent les erreurs de mesures avec les hypothèses $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\delta] = 0$ et il est supposé que ε et δ sont non-correlées avec ζ et ξ. Le modèle est ainsi caractérisé par un ensemble de paramètres θ qui intègre les matrices $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Lambda_y}$, $\mathbf{\Lambda_x}$ ainsi que les variances d'erreurs ζ, ξ, ε et δ.

Application. Pour notre exemple illustratif, nous avons $p = 4$ variables exogènes (PHO, PAK, GLU et Bmc) et $q = 2$ variables endogènes (BSr et Ntl). Les 4 mesures $PHOS$, PAK, GLU et Bmc sont caractéristiques des propriétés du sol, la mesure BSr est obtenue au niveau des racines tandis que Ntl est un indicateur des propriétés du couvert. Notre système d'étude peut donc se décomposer en 3 compartiments modélisés par 3 variables latentes : $n = 1$ variable latente exogène (sol) et $m = 2$ variables latentes endogènes (rcn pour racine et arn pour aerien). En faisant l'hypothèse que les propriétés du sol influence l'activité des racines qui elle même va influencer les propriétés du couvert. Avec $\mathbf{x} = [PHO, PAK, GLU, Bmc]$, $\mathbf{y} = [Bsr, Ntl]$, $\xi = [sol]$ et $\eta = [rcn, arn]$ nous avons :

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{1,1} \\ 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & b_{2,2} \end{bmatrix}, \mathbf{\Lambda_x} = \begin{bmatrix} \lambda^x_{1,1} \\ \lambda^x_{2,1} \\ \lambda^x_{3,1} \\ \lambda^x_{4,1} \end{bmatrix}, \mathbf{\Lambda_y} = \begin{bmatrix} \lambda^y_{1,1} & 0 \\ 0 & \lambda^y_{2,2} \end{bmatrix}.$$

Le schéma conceptuel sous la forme d'un graphe acyclique orienté est représenté Figure 1 (à gauche).

2.2 Estimation et godness-of-fit

L'estimation d'un modèle d'équations structurelles se fait classiquement par l'optimisation d'une fonction d'ajustement, parmi lesquelles la fonction d'ajustement du maximum de vraisemblance est la plus utilisée (Rosseel, 2012). En s'appuyant sur l'hypothèse que la matrice de covariance des variables observées suit un loi de Wishart, la fonction d'ajustement s'écrit :

$$F_{ML}(\Sigma(\theta)) = \log|\Sigma(\theta)| + \operatorname{tr}\left(S\Sigma^{-1}(\theta)\right) - \log|S| - (p+q) \tag{4}$$

où S est la matrice de variance-covariance observée et $\Sigma(\theta)$ celle induite par le modèle. L'adéquation du modèle aux données consiste à comparer le modèle estimé au modèle saturé pour lequel $\Sigma(\theta) = S$ et un nombre de degrés de liberté égal à 0, comme par exemple en utilisant un test de rapport de vraisemblance (Rosseel, 2012).

Application. La Figure 1 présente le modèle agroécologique composé des 6 variables manifestes et des 3 variables latentes. Pour tester l'adéquation de ce modèle aux données, il sera confronté au modèle saturé qui peut être décrit par un modèle sans variable latente pour lequel toutes les variances et les covariances entre variables observées sont des paramètres à estimer. Notre modèle agroécologique (modèle testé) est confirmé par les données avec une p-valeur du test du rapport de vraisemblance égal à 0.2381.

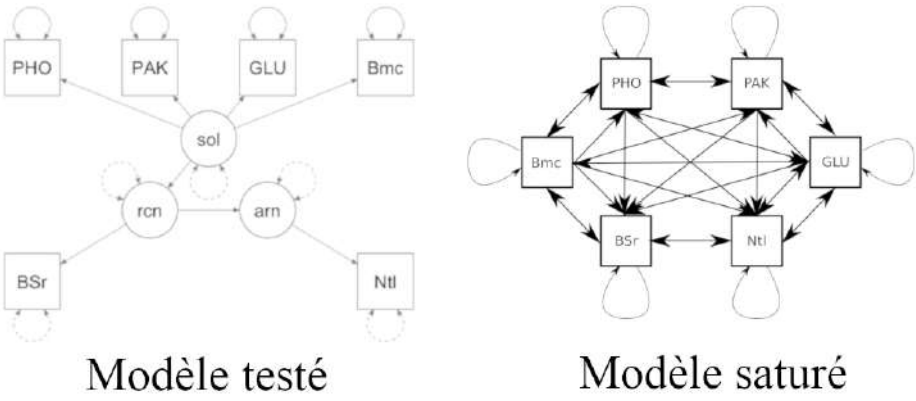


Figure 1: La figure de gauche correspond à une visualisation par graphe orienté acyclique du modèle testé. La figure de droite représente le modèle saturé qui correspond à un graphe complet.

3 Sélection de variables

La fonction objective présentée à l'équation (4) permet d'estimer et de comparer des modèles SEM construits pour un ensemble fixe de variables observées. Dans cette section, nous nous intéressons à quantifier l'impact de la présence ou de l'absence d'une variable observée, $x^\star$, dans le système causal. Soit l'ensemble des variables observées $(\mathbf{y}, \mathbf{x}) = (\mathbf{y_1}, \mathbf{x_1}, x^\star)$, $\Sigma(\theta)$ peut se réécrire de la façon suivante :

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_1(\theta) & u(\theta) \\ u(\theta)' & \mathbb{V}_\theta(x^\star) \end{bmatrix}$$

où $\Sigma_1(\theta)$ est le bloc de la matrice de covariance, obtenu à partir de l'estimation de θ, pour le sous-ensemble de variables observées $(\mathbf{y_1}, \mathbf{x_1})$. $u(\theta)$ est l'ensemble des covariances entre $x^\star$ et $(\mathbf{y_1}, \mathbf{x_1})$ et $\mathbb{V}_\theta(x^\star)$ est la variance de $x^\star$. Considérons également $\Sigma_0(\theta_0)$, la matrice de variance-covariance des variables observées $(\mathbf{y_1}, \mathbf{x_1})$, induite par le modèle structural estimé uniquement sur $(\mathbf{y_1}, \mathbf{x_1})$. Une illustration est proposée à la Figure 2.

Pour déterminer l'importance de la variable x dans le système causal, nous définissons la statistique D comme la différence des fonctions d'ajustements :

$$D = F_{ML}(\Sigma_1(\theta)) - F_{ML}(\Sigma_0(\theta_0)). \tag{5}$$

La significativité de D est testée par permutation des valeurs de $x^\star$. Pour une permutation $\mathcal{P}$, les paramètres du modèle $\theta_\mathcal{P}$ sont estimés pour calculer une mesure d'ajustement $F_{ML}(\Sigma_1(\theta_\mathcal{P}))$.

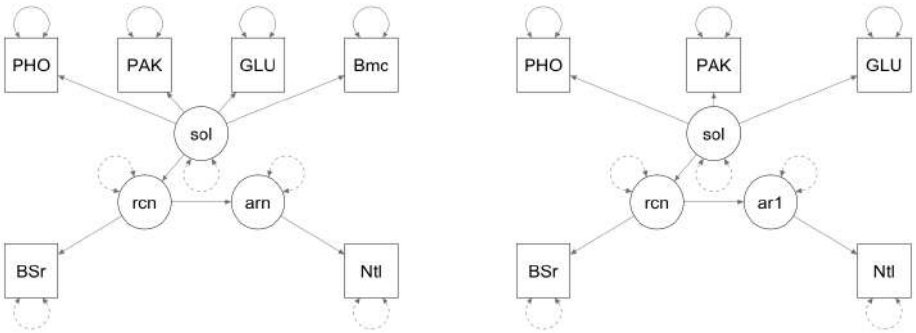


Figure 2: La figure de gauche correspond au modèle impliquant la variable *Bmc* et la figure de droite au modèle sans la variable *Bmc*.

Application. Dans notre application, nous cherchons à tester si les données biomoléculaires (la variable *Bmc*) sont nécessaires à l'estimation du système causal. La valeur observée pour D est égale à 0.06991. A partir de 1000 permutations de la variable *Bmc*, la p-valeur obtenue est égale à 0.55.

4 Conclusion

Dans cette étude, nous avons proposé une méthode de sélection de variables pour des modèles d'équations structurelles. Cette méthode permet d'étudier la composition de l'ensemble des variables observées nécessaires à l'estimation d'une structure causale. L'application de notre méthode dans un contexte agroécologique a permis de mettre en évidence la robustesse du modèle causal en l'absence de la mesure de biomasse microbienne. L'utilisation de notre modèle peut permettre de reconsidérer les importances relatives des variables du processus et ainsi éviter la multiplication des mesures. Ainsi, nos travaux offrent des perspectives quant à l'utilisation des modèles d'équations structurelles dans la planification expérimentale.

Bibliographie

Benkeser, D. and Chambaz, A. and Van der Laan, Mark J. (2020) Causality: a special issue of Journal de la Société Française de Statistique, *Journal de la Société Française de Statistique*, 161, pp. 1-3

Bollen, K. A. (1989). *Structural equations with latent variables*, Wiley.

Cai, X., Bazerque, J. A. and Giannakis, G. B. (2013) Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations, *PLoS Computational Biology*, 9.

Jacobucci, R., Grimm, K. J. and McArdle, J.J. (2016) Regularized Structural Equation Modeling, *Structural Equation Modeling*. 23, pp. 555–566.

Petersen, M. L. and van der Laan, M. J. (2014) Causal models and learning from data: integrating causal modeling and statistical estimation, *Epidemiology*, 25, pp 418-426.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling, *Journal of Statistical Software*, 48, pp. 1-36.

Shipley, B. (2016), *Cause and correlation in Biology*, Cambridge University Press, 2nd Edition.

Estimation of the Cure Rate for Distributions in the Gumbel Maximum Domain of Attraction Under Insufficient Follow-up.

Mikael Escobar-Bach[1] Ross Maller[2]
Ingrid Van Keilegom[3] Muzhi Zhao[2]

[1]*LAREMA, Université d'Angers, France.*
[2]*ORSTAT, KU Leuven, Belgium.*
[3]*RSFAS, Australian National University, Australia.*

Résumé. Estimer le taux d'immunité à partir des données de survie peut se réléver difficile lorsque ces dernières comprennent des sujets *immunisés* qui n'expérimenterons jamais l'événement d'intérêt. Dans un certain sens, on peut s'attendre à ce que tout estimateur proposé ne donne de bons résultats que lorsque la période de suivi est suffisante. Dans le contexte d'un suivi insuffisant, Escobar-Bach et Van Keilegom (2019) proposent un estimateur non-paramétrique qui incorpore un terme d'ajustement améliorant l'estimation, valable sous l'hypothèse que la distribution de survie des sujets *susceptibles* à l'événement appartient au domaine maximum d'attraction de Fréchet. Avec ce travail, nous choisissons d'étendre leur méthode aux distributions appartenant au domaine maximum d'attraction de Gumbel, qui représente lui aussi une large classe de distributions utiles en survie. A l'instar de Escobar-Bach et Van Keilegom (2019), nous utilisons des techniques d'extrapolation de la théorie des valeurs extrêmes pour construire un estimateur non-paramétrique consistant et améliorant l'estimation du taux d'immunité. Nous présentons également les propriétés asymptotiques de l'estimateur et proposons une application sur des données de survie de patients à différents stades du cancer du sein.

Mots clés : Analyse de survie, modèle de mélange avec *cure*, théorie des valeurs extrêmes, suivi insuffisant.

Abstract. Estimating the cured proportion from survival data which may include observations on cured subjects, that is, those who never experience the event of interest, is an important task in practice. Any proposed estimator can only be expected to perform well when the follow-up period is sufficient, in some sense. When follow-up is not sufficient, a nonparametric estimator proposed by Escobar-Bach and Van Keilegom (2019) incorporates an adjustment which ameliorates the problem under the assumption that the survival distribution of those susceptible to the event belongs to the Fréchet maximum domain of attraction. In the present work we extend their method to distributions belonging to the Gumbel maximum domain of attraction, a significant extension since many commonly used lifetime distributions have this property. Like Escobar-Bach and Van Keilegom (2019) we use extrapolation techniques from extreme value theory to derive

a nonparametric estimator of the cure proportion which is consistent and approximately normally distributed under certain assumptions, and performs well in simulation studies. We illustrate with an application to survival data where patients with differing stages of breast cancer have varying degrees of follow-up.

Key words : Survival analysis, mixture cure model, extreme value theory, insufficient follow-up.

1 Contexte

Modèle de mélange immunisé

On considère une expérimentation où chaque individu est associé à un temps de survie aléatoire $T \in [0, +\infty]$ et un censeur aléatoire $C \in \mathbb{R}$:

- si $T < +\infty$, le sujet fait partie des *susceptibles* et est censuré lorsque $T > C$.
- si $T = +\infty$, le sujet est dit *immunisé* et est automatiquement censuré.

D'un point de vue de l'utilisateur, l'observation est réduite à $Y := \min(T, C)$ avec pour indicateur de censure $\delta := \mathbb{1}_{\{T \leq C\}}$. Le modèle de mélange immunisé est alors caractérisé par les fonctions de distribution

$$F_c(t) := \mathbb{P}(C \leq t) \quad \text{et} \quad F(t) := \mathbb{P}(T \leq t) = pF_0(t) \tag{1}$$

où $p := \mathbb{P}(T < +\infty)$, T et C sont indépendants et F_0 définit la fonction de distribution des susceptibles. On introduit également les points terminaux respectifs des fonctions F_0 et F_c par τ_0 et τ_c.

Estimation du taux d'immunité et problème de suivi

Dans le but d'estimer le taux d'immunité $1 - p$, la plupart des études reposent sur l'estimateur de Kaplan et Meier (1958). Par définition, il s'agit d'une estimateur produit pour la fonction F, qui à l'avantage de tirer parti des données censurées et non-censurées. Afin de l'introduire, on note $\{(Y_i, \delta_i)\}_{1 \leq i \leq n}$ un échantillon *i.i.d.* de (1) et on introduit les notations $Y_{(i)}$, pour la i-ième statistique d'ordre, et $\delta_{(i)}$, pour l'indicateur de censure correspondant. L'estimateur de Kaplan-Meier est ensuite donné par

$$\widehat{F}_n(t) := 1 - \prod_{Y_{(i)} \leq t} \left(1 - \frac{\delta_{(i)}}{n - 1 + i}\right); \quad t \in \mathbb{R}$$

où tout produit sur l'ensemble vide retourne la valeur 1. La valeur p est alors approchée via (1), à l'aide de l'estimateur

$$\widehat{p}_n := \widehat{F}_n(\widehat{\tau}_n)$$

où le maximum des données de survie est simplement noté $t_{(n)} := Y_{(n)}$. D'après Maller et Zhou (1992), ce dernier est un estimateur consitant de p uniquement pour des cas de suivi suffisant :

Théorème 1.1 *On suppose que* $0 < p \leq 1$ *et* F *continue à* $\tau_G := \tau_0 \wedge \tau_c$ *lorsque* $\tau_G < +\infty$*, alors on obtient que*

$$\widehat{p}_n \xrightarrow{p.s} p \text{ lorsque } n \to +\infty \quad \text{si et seulement si} \quad \tau_0 \leq \tau_c.$$

En conséquence, la consistence de l'estimateur $\widehat{p}_n$ requiert la condition nécessaire et suffisante qu'aucun susceptible ne puisse survive plus longtemps que le plus grand des censeurs. Cette condition est communément nommée *suivi suffisant*, et représente un paradigme usuel en analyse de survie. En pratique, il est compliqué de vérifier cette dernière et des difficultés peuvent apparaître pour des expérimentations avec un cours temps d'étude. En particulier, l'estimateur $\widehat{p}_n$ sous-estime p lorsque $\tau_c < \tau_0$, condition dite de *suivi insuffisant*, où $\tau_c < +\infty$ et τ_0 est possiblement infini.

2 Estimation dans le cas de suivi insuffisant

Domaine d'attraction pour la distribution des susceptibles

Dans ce projet, nous proposons de palier au problème de suivi insuffisant en se basant sur la connaissance du comportement limite de F proche de τ_0. Ce type de raisonnement est propre à la théorie des valeurs extrêmes, en particulier lorsqu'il s'agit d'extrapoler des probabilités d'évènements rares. Nous supposons donc que F_0 appartient au domaine d'attraction maximum d'une distribution des valeurs extrêmes, c'est-à-dire, on suppose qu'il existe un paramètre $\gamma \in \mathbb{R}$ tel que pour tout $y > 0$

$$\lim_{t \to \tau_0} \frac{1 - F_0(t + y\ell(t))}{1 - F_0(t)} = G_\gamma(y) := \begin{cases} (1+\gamma y)^{-1/\gamma}, & \text{si } \gamma \neq 0, \\ \exp(-y), & \text{si } \gamma = 0, \end{cases} \tag{2}$$

avec

$$\ell(t) = \begin{cases} \gamma t, & \text{si } \gamma > 0, \\ -\gamma(\tau_0 - t), & \text{si } \gamma < 0, \\ \displaystyle\int_t^{\tau_0} 1 - F_0(x)dx/(1 - F_0(t)), & \text{si } \gamma = 0. \end{cases} \tag{3}$$

Le paramètre γ est *l'indice des valeurs extrêmes* et caractérise le comportement de queue de F_0. Dans cette étude, nous nous intéressons au cas où $\gamma = 0$. L'estimation de ℓ dans (3) joue alors un rôle majeur dans l'extrapolation de F_0 bien qu'elle ne soit pas directement accessible dans notre contexte, car dépendante de F_0 sur tout son support.

Construction de l'estimateur

L'hypothèse (2) nous permet d'atteindre des valeurs normalement hors de portée à cause de la contrainte de censure particulière. Nous proposons donc d'étendre $\widehat{p}_n$ en lui sommant une correction positive dépendante de τ_G. Formellement, nous estimons dans un premier temps la fonction ℓ en $\tau_G - \varepsilon$ avec $\varepsilon > 0$ suffisamment petit, d'où

$$\ell(\tau_G - \varepsilon) \simeq -\frac{\varepsilon}{2\log(1/Z(\varepsilon) - 1)} \quad \text{avec} \quad Z(\varepsilon) = \frac{F\left(\tau_G - \varepsilon/2\right) - F\left(\tau_G - \varepsilon\right)}{F\left(\tau_G\right) - F\left(\tau_G - \varepsilon\right)}.$$

En utilisant de nouveau (2), on obtient alors une approximation de p via

$$p \simeq F(\tau_G - \varepsilon) + \big(F(\tau_G) - F(\tau_G - \varepsilon)\big)\big(1 - \exp\left(-\varepsilon/l(\tau_G - \varepsilon)\right)\big)^{-1}.$$

En remplaçant τ_G, ℓ et F par leurs estimateurs respectifs, on obtient finalement

$$\begin{aligned} \hat{p}_G(n,\varepsilon) &= \hat{F}_n(t_{(n)} - \varepsilon) + \big(\hat{F}_n(t_{(n)}) - \hat{F}_n(t_{(n)} - \varepsilon)\big)\big(1 - \exp\big(-\varepsilon/\hat{l}(t_{(n)} - \varepsilon)\big)\big)^{-1} \\ &:= \hat{F}_n(t_{(n)} - \varepsilon) + \hat{C}(t_{(n)}, \varepsilon) \end{aligned}$$

que l'on peut réécrire après simplification

$$\hat{p}_G(n,\varepsilon) = \hat{F}_n(t_{(n)} - \varepsilon) + \frac{\big(\hat{F}_n(t_{(n)} - \varepsilon/2) - \hat{F}_n(t_{(n)} - \varepsilon)\big)^2}{2\hat{F}_n\left(t_{(n)} - \varepsilon/2\right) - \hat{F}_n(t_{(n)} - \varepsilon) - \hat{F}_n(t_{(n)})}.$$

Résultat asymptotique

Afin d'obtenir des résultats asymptotiques sur nos estimateurs, nous travaillerons dans une classe de distributions légèrement plus petite que le domaine d'attraction maximum de Gumbel complet, avec des distributions satisfaisant la condition de *Von Mises*. Cette condition est suffisante mais non nécessaire pour appartenir au domaine d'attraction, bien que en pratique, elle impose très peu de restriction.

Hypothèse 2.1 *On suppose que F_0 admet une dérivée seconde finie et négative F_0'' au voisinage de τ_0 avec la propriété limite suivante*

$$\lim_{t \uparrow \tau_0} \frac{F_0''(t)\left(1 - F_0(t)\right)}{\left(F_0'(t)\right)^2} = -1.$$

Sous cette dernière hypothèse, nous pouvons alors énoncer les principaux résultats de cette étude.

Théorème 2.1 (Consistence de $\hat{p}_G(n,\varepsilon)$) *Dans le cadre d'un modèle de censure droite i.i.d. avec $\tau_G < \tau_0$ et sous l'hypothèse 2.1, quelque soit $\varepsilon > 0$, lorsque $n \to \infty$,*

$$\hat{C}(t_{(n)},\varepsilon) \xrightarrow{p} C(\tau_G,\varepsilon) := \frac{\left(F\left(\tau_G - \varepsilon/2\right) - F\left(\tau_G - \varepsilon\right)\right)^2}{2F\left(\tau_G - \varepsilon/2\right) - F\left(\tau_G - \varepsilon\right) - F\left(\tau_G\right)},$$

$$C(\tau_G,\varepsilon) \to -p\frac{F_0'\left(\tau_G\right)^2}{F_0''\left(\tau_G\right)} = C(\tau_G).$$

Lorsque $n \to \infty$, $\varepsilon \to 0$ et $\tau_G \to \tau_{F_0}$, l'estimateur $\hat{p}_G(n,\varepsilon)$ est consistant au sens qu'il converge vers $F(\tau_G) + C(\tau_G)$, où de plus,

$$\lim_{\tau_G \to \tau_{F_0}} F(\tau_G) + C(\tau_G) = p.$$

Finalement, le théorème suivant garantie la normalité asymptotique de $\hat{p}_G(n,\varepsilon)$.

Théorème 2.2 (Normalité asymptotique de $\hat{p}_G(n,\varepsilon)$) *Dans le cadre d'un modèle de censure droite i.i.d. avec $\tau_G < \tau_0$ et sous l'hypothèse 2.1, on suppose que $\lim_{n\to\infty} n\bar{G}(\tau_H - \delta/\sqrt{n}) = \infty$ pour $\delta > 0$. On suppose également que l'hyptohèse d'intégrabilité (3.61) de Maller and Zhou (1996) est vérifiée. Alors $\hat{p}_G(n,\varepsilon)$ est asymptotiquement normal lorsque $n \to \infty$ quelque soit $\varepsilon > 0$, et la distribution limite est normal lorsque $\varepsilon \to 0$.*

Simulation et application réelle

Ces résultats seront illustrés numériquement lors de l'exposé. Dans un premier temps, nous comparerons $\widehat{p}_y$ et l'estimateur de Kaplan-Meier pour différentes distributions simulées. Dans un second temps, nous montrerons leur utilisation en pratique avec des suivis sur le cancer du sein issus de la base de données Surveillance, Epidemiology and End Results (SEER). L'expérimentation portera sur l'estimation des taux d'immunité aux Etats-Unis pour des femmes atteintes du cancer de la phase I à IV.

Bibliographie

[1] E. L. Kaplan and Paul Meier. (1958), Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

[2] R. A. Maller and X. Zhou. (1992), Estimating the proportion of immunes in a censored sample. *Biometrika*, 79(4):731–739.

[3] R. A. Maller and X. Zhou. (1996), *Survival analysis with long-term survivors*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

[4] Escobar-Bach M. and Van Keilegom I. (2019), *Non-parametric cure rate estimation under insufficient follow-up by using extremes* J. R. Statist. Soc. B. 81(5)861–880.

Detecting spatial clusters in functional data: new scan statistic approaches

Camille Frévent [1] & Mohamed-Salem Ahmed [1] & Matthieu Marbac [2] & Michaël Genin [1]

[1] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.* camille.frevent@univ-lille.fr, mohamed-salem.ahmed@univ-lille.fr, michael.genin@univ-lille.fr

[2] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France.* matthieu.marbac-lourdelle2@ensai.fr

Résumé. Nous avons développé deux statistiques de scan pour détecter des clusters sur des données fonctionnelles indexées dans l'espace. La première méthode est basée sur une adaptation de l'ANOVA fonctionnelle et la deuxième est basée sur une statistique de scan spatiale *distribution-free* pour données univariées. Dans une simulation, la deuxième méthode présente toujours de meilleures performances qu'une statistique de scan non paramétrique pour données fonctionnelles, et l'adaptation de l'ANOVA présente également de meilleures performances pour des données normales. Nos approches détectent de plus petits clusters que la méthode non paramétrique. Enfin nous avons appliqué nos statistiques de scan spatiales sur des données fonctionnelles pour détecter des clusters spatiaux de taux de chômage anormalement élevés ou faibles en France sur la période 1998-2013.
Mots-clés. Détection de clusters, données fonctionnelles, statistiques de scan spatiales

Abstract. We have developed two scan statistics for detecting clusters of functional data indexed in space. The first method is based on an adaptation of a functional analysis of variance and the second one is based on a distribution-free spatial scan statistic for univariate data. In a simulation study, the distribution-free method always performed better than a nonparametric functional scan statistic, and the adaptation of the ANOVA also performed better for data with a normal distribution. Our methods can detect smaller spatial clusters than the nonparametric method. Lastly, we used our scan statistics for functional data to search for spatial clusters of abnormal unemployment rates in France over the period 1998-2013 (divided into quarters).
Keywords. Cluster detection, functional data, spatial scan statistics

1 Introduction

Les méthodes de détection de clusters spatiaux ont été longuement étudiées. Le but est de développer des outils capables de détecter l'aggrégation de sites qui se comportent

différemment. Les statistiques de scan spatiales permettent de détecter des clusters spatiaux sans information *a priori* quant à leur localisation. Elles ont été principalement proposées par Kulldorff et Nagarwalla (1995) et Kulldorff (1997) dans le cas de modèles de Poisson et de Bernouilli. D'autres statistiques de scan ont ensuite été proposées pour d'autres modèles de distribution (Kulldorff *et al.* (2009), Cucala *et al.* (2017)).
Aujourd'hui les données sont de plus en plus mesurées en temps continu ou quasi-continu. Cela a conduit au développement de méthodes d'analyse de données fonctionnelles (Ramsay et Silverman (2005)). Dans le domaine des statistiques de scan spatiales, l'application d'une méthode univariée entraînerait une très grande perte d'information. Une méthode multivariée, considérant chaque temps d'observation comme une variable sera confrontée à des problèmes de grandes dimensions et de grandes corrélations. Récemment Smida *et al.* (2020) a développé une méthode non paramétrique basée sur un test de Wilcoxon-Mann-Whitney fonctionnel. Aucune statistique de scan spatial paramétrique n'a été développée pour des données fonctionnelles. Puisqu'un test d'ANOVA pour données fonctionnelles a été décrit par Cuevas *et al.* (2004), nous allons développer une statistique de scan basée sur ce test. Ensuite puisque ces dernières années des tests statistiques ont été développés pour des données en grande dimension, en résumant l'information après le calcul d'une statistique pour chaque composante (Lin *et al.* (2021)), nous allons proposer une statistique de scan basée sur la combinaison de cette approche et de la statistique de scan *distribution-free* pour données univariées proposée par Cucala (2014).

Nous décrivons ici la statistique de scan spatial paramétrique pour données fonctionnelles basée sur une ANOVA fonctionnelle ainsi qu'une autre méthode basée sur la statistique de scan *distribution-free*. La méthodologie est présentée en Section 2. Dans la section 3, les performances des méthodes sont étudiées et comparées avec celles de Smida, *et al.* (2020) et les méthodes sont appliquées sur des données réelles. Enfin, une discussion est présentée en Section 4.

2 Méthodologie

Soit $s_1, \ldots, s_n$, n sites d'un domaine d'observation $S \subset \mathbb{R}^2$ et $X_1, \ldots, X_n$ les observations de X dans $s_1, \ldots, s_n$. Ici, $\{X(t),\ t \in \mathcal{T}\}$ est un processus stochastique à valeurs réelles, avec $\mathcal{T}$ un intervalle de $\mathbb{R}$. A partir de maintenant les observations sont considérées indépendantes, ce qui est une hypothèse classique en statistique de scan.
Le but est de détecter des clusters spatiaux et de tester leur significativité. On cherche donc à tester $\mathcal{H}_0$ (l'absence de cluster) contre une hypothèse alternative $\mathcal{H}_1$ (la présence d'au moins un cluster $w \subset S$ de valeurs anormales pour X).

Cressie, N. (1977) définit une statistique de scan spatial comme le maximum d'un indice

de concentration sur un ensemble de clusters potentiels $\mathcal{W}$. Dans la suite sans perte de généralité, $\mathcal{W}$ est un ensemble de clusters circulaires contenant entre 1 et 50% des sites.

2.1 Statistique de scan paramétrique pour données fonctionnelles

On suppose que le processus X est à valeurs dans l'espace $L^2(\mathcal{T}, \mathbb{R})$ des fonctions réelles de carré intégrable sur $\mathcal{T}$.
Cuevas *et al.* (2004) and Górecki et Smaga (2015) ont adapté la F-statistique de l'ANOVA pour les processus L^2. Sans perte de généralité, en considérant deux échantillons indépendants provenant de deux processus L^2 X_{g_1} and X_{g_2} dans deux groupes g_1 et g_2, le test compare les fonctions moyennes μ_{g_1} and μ_{g_2} where $\mu_{g_i}(t) = \mathbb{E}[X_{g_i}(t)]$, $i = 1, 2$.
Alors, pour la détection de cluster, $\mathcal{H}_0$ peut être définie par : $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^c} = \mu_S$, où μ_w, μ_{w^c} et μ_S sont les fonctions moyennes dans w, à l'extérieur de w and dans S, respectivement. L'hypothèse alternative associée au cluster potentiel w $\mathcal{H}_1^{(w)}$ se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$. Ensuite l'ANOVA fonctionnelle peut-être utilisée pour comparer les fonctions moyennes dans w et dans w^c en utilisant la statistique suivante :

$$F_n^{(w)} = \frac{|w|\ ||\bar{X}_w - \bar{X}||_2^2 + |w^c|\ ||\bar{X}_{w^c} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j, s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j, s_j \in w^c} ||X_j - \bar{X}_{w^c}||_2^2\right]}, \tag{1}$$

où $\bar{X}_g(t)$ sont les estimateurs empiriques de μ_g ($g \in \{w, w^c\}$), $\bar{X}(t)$ est l'estimateur empirique de μ_S et $||x||_2^2 = \displaystyle\int_{\mathcal{T}} x^2(t)\ \mathrm{d}t$.
Alors, $F_n^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$, ce qui amène à la définition suivante de la statistique de scan spatial paramétrique pour données fonctionnelles (PFSS) : $\Lambda_{\text{PFSS}} = \max\limits_{w \in \mathcal{W}} F_n^{(w)}$. Le cluster potentiel pour lequel ce maximum est atteint est appelé "cluster le plus probable" (*most likely cluster* (MLC)) est donc $\text{MLC} = \arg\max\limits_{w \in \mathcal{W}} F_n^{(w)}$.

2.2 Statistique de scan *distribution-free* pour données fonctionnelles

Ici nous proposons de combiner la statistique de scan *distribution-free* pour données univariées proposée par Cucala (2014) et la statistique "max" de Lin *et al.* (2021). Brièvement ces derniers proposent une nouvelle approche au problème de l'ANOVA fonctionnelle en maximisant une statistique au cours du temps.

Nous supposons que pour chaque temps t, $\mathbb{V}[X_i(t)] = \sigma^2(t)\ \ \forall i \in [\![1; n]\!]$. Alors pour chaque t, l'indice de concentration proposé par Cucala (2014) pour tester $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) =$

$\mu_{w^c}(t) = \mu_S(t)$ est

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^c}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)]}}, \text{ où } \hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)] = \widehat{\sigma^2}(t) \left[\frac{1}{|w|} + \frac{1}{|w^c|}\right],$$

avec $\widehat{\sigma^2}(t) = \frac{1}{n-2}\left[\sum_{i,s_i \in w} (X_i(t) - \bar{X}_w(t))^2 + \sum_{i,s_i \in w^c} (X_i(t) - \bar{X}_{w^c}(t))^2\right]$.

Maintenant l'idée est de globaliser l'information en maximisant la quantité précédente au cours du temps pour chaque cluster potentiel w (Lin *et al.* (2021)) : $I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t)$.

Pour la détection de cluster, comme pour le PFSS, $\mathcal{H}_0$ peut être définie par $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^c} = \mu_S$. Et l'hypothèse alternative $\mathcal{H}_1^{(w)}$ associée au cluster potentiel w se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$.

$I^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$ pour obtenir la statistique de scan spatial *distribution-free* pour données fonctionnelles (DFFSS) : $\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} I^{(w)}$. Et pour cette méthode le *most likely cluster* est défini par $\text{MLC} = \arg\max_{w \in \mathcal{W}} I^{(w)}$.

2.3 Significativité du MLC

Une fois le MLC détecté, sa significativité doit être évaluée. Pour cela nous générons des données par permutation des données de départ (*random labelling*) et nous en déduisons une estimation de la p-valeur (Dwass (1957)). Enfin le MLC est considéré significatif si la p-valeur associée est inférieure à l'erreur de type I.

3 Résultats

3.1 Etude d'une simulation

Nous avons comparé (i) la statistique de scan spatial paramétrique pour données fonctionnelles (PFSS) Λ_{PFSS}, (ii) la statistique de scan spatial *distribution-free* pour données fonctionnelles (DFFSS) Λ_{DFFSS} et (iii) la statistique de scan spatial non paramétrique pour données fonctionnelles (NPFSS) Λ_{NPFSS} développée par Smida *et al.* (2020).

3.1.1 Plan de la simulation

Les sites considérés sont les 94 départements de France métropolitaine, localisés par leur centre administratif. Un vrai cluster composé des 8 départements d'Île-de-France est défini pour chacune des bases de données simulées.

Les X_i sont générés avec le modèle suivant :

$$\text{pour } i \in [\![1;94]\!],\ X_i(t) = \sin\left[2\pi t^2\right]^5 + \Delta(t)\mathbb{1}_{s_i \in w} + \varepsilon_i(t),\ t \in [0;1].$$

où $\varepsilon_i(t) = \sum_{k=1}^{7} \sqrt{1.5 \times 0.2^k}(v_{i,1,k} - v_{i,2,k})\Psi_k(t)$ et $\Psi_k(t) = \begin{cases} 1 & \text{si } k = 1 \\ \sqrt{2}\sin[k\pi t] & \text{si } k \text{ pair} \\ \sqrt{2}\cos[(k-1)\pi t] & \text{si } k \text{ impair, } k > 1 \end{cases}$.
Trois distributions pour les $v_{i,j,k}, j = 1, 2$ sont considérées : $v_{i,j,k} \sim \mathcal{N}(0,1)$, $v_{i,j,k} \sim t(4)/\sqrt{2}$ et $v_{i,j,k} \sim (\chi^2(4) - 4)/(2\sqrt{2})$.
Trois types de clusters sont simulés avec $\Delta_1(t) = \alpha t$, $\Delta_2(t) = \alpha t(1-t)$ et $\Delta_3(t) = \alpha \exp[-100(t-0.5)^2]/3$, où α est un paramètre controllant l'intensité du cluster.

Pour chaque distribution, chaque Δ, et différentes valeurs de α, nous avons simulé 1000 bases de données. La p-valeur associée à chaque MLC est estimée en générant 999 permutations des données et nous considérons une erreur de type I de 0.05.

3.1.2 Résultats de la simulation

Le DFFSS présente de meilleures performances que les deux autres méthodes, surtout dans le cas du shift local Δ_3. Le PFSS et le NPFSS présentent des performances similaires dans le cas gaussien mais les performances du PFSS diminuent quand on s'éloigne de la normalité. Le NPFSS a tendance à détecter des clusters plus grands que le PFSS et le DFFSS.

3.2 Application sur données réelles

Nous avons considéré les données de taux de chômage en France métropolitaine pour chaque trimestre entre 1998 et 2013.
Le PFSS et le DFFSS détectent un même MLC significatif composé de 7 départements dans le sud-est de la France. Dans ce cluster les courbes des taux de chômage sont toutes au-dessus de la courbe du chômage moyen en France. En revanche le NPFSS détecte un MLC significatif de 40 départements dans le centre de la France. La plupart des courbes dans ce cluster sont en-dessous de la courbe de chômage moyen, cependant certaines sont au-dessus ce qui traduit une hétérogénéité dans le cluster.

4 Discussion

Même si le DFFSS et le PFSS permettent de traiter uniquement des données fonctionnelles univariées, le cas fonctionnel multivarié devrait être étudié également. Par exemple on pourrait considérer des données collectées par des capteurs situés dans différentes zones géographiques et mesurant plusieurs polluants de l'air au cours du temps. Le PFSS

peut être étendu au cadre multivarié en adaptant l'ANOVA multivariée pour des données fonctionnelles comme suggéré par Górecki et Smaga (2017). Le test de Wilcoxon-Mann-Whitney fonctionnel (Chakraborty et Chaudhuri (2014)) est aussi adaptable au cadre multivarié, en prenant un produit scalaire judicieux pour calculer la fonction signe.

Bibliographie

Chakraborty, A. et Chaudhuri, P. (2014). A Wilcoxon-Mann-Whitney type test for infinite dimensional data, *Biometrika*, 102.

Cressie, N. (1977). On Some Properties of the Scan Statistic on the Circle and the Line, *Journal of Applied Probability*, 14, pp. 272-283.

Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes, *Spatial Statistics*, 10, pp. 117–125.

Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A Multivariate Gaussian scan statistic for spatial data, *Spatial Statistics*, 21, pp. 66-74.

Cuevas, A., Febrero-Bande, M. et Fraiman, R. (2004). An ANOVA test for functional data, *Computational Statistics & Data Analysis*, 47, pp. 111-122.

Dwass, M. (1957). Modified Randomization Tests for Nonparametric Hypotheses, *Annals of Mathematical Statistics*, 28, pp. 181-187.

Górecki, T. et Smaga, Ł. (2015). A comparison of tests for the one-way ANOVA problem for functional data, *Computational Statistics*, 30, pp. 987-1010.

Górecki, T. et Smaga, Ł. (2017). Multivariate analysis of variance for functional data, *Journal of Applied Statistics*, 44(12), pp. 2172-2189.

Kulldorff, M. et Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, 14(8), pp. 799-810.

Kulldorff, M. (1997). A Spatial Scan Statistic, *Communications in Statistics - Theory and Methods*, 26, pp. 1481-1496.

Kulldorff, M. et Huang, L. and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model, *Int J Health Geogr*, 8(58).

Lin, Z., M. E. Lopes et H.-G. Müller (2021). High-dimensional manova via bootstrapping and its application to functional and sparse count data.

Ramsay, JO et Silverman, BW (2005). Functional Data Analysis *Springer*.

Smida, Z., Cucala, L. et Gannoun, A. (2020). A nonparametric spatial scan statistic for functional data, *hal-02908496*.

Approche bayésienne à l'estimation de la zone du langage chez des patients ayant eu un AVC

Mame Diarra Fall [1], Nicolas Dobigeon [2] & Pascal Auzou [3]

[1] *Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS.*
Email : diarra.fall@univ-orleans.fr
[2] *Université de Toulouse, IRIT/INP-ENSEEIHT, 31071 Toulouse, France.*
Email : dobigeon@n7.fr
[3] *Centre Hospitalier Régional d'Orléans, Service de Neurologie.*
Email : pascal.auzou@chr-orleans.fr

Résumé. L'étude de patients cérébro-lésés est une approche importante pour examiner des relations structure-fonction en utilisant des techniques de neuro-imagerie. La méthode VLSM (Voxel-based Lesion Symptom Mapping, [BWS+03]) a été largement utilisée pour identifier des régions cérébrales impliquées dans différentes fonctions cognitives. Elle consiste à utiliser des images recalées de nombreux patients puis, en effectuant des tests statistiques voxel par voxel, d'inférer sur les régions concernées par une certaine pathologie (par exemple, les troubles du langage). La méthode VLSM est cependant basée sur l'utilisation du test de Student pour lequel l'hypothèse de normalité n'est pas toujours vérifiée et nécessite donc souvent un nombre important de patients. Afin de répondre à cette limitation, nous avions proposé dans [FLP18] un test bayésien non paramétrique utilisant les arbres de Pólya. Cependant, cette méthode, tout comme celles de la littérature, traite un voxel indépendamment de ses voisins. Nous proposons ici d'utiliser une approche totalement différente. On aborde le problème, non plus comme celui d'un test statistique, mais plutôt comme celui d'une estimation statistique. L'objectif est d'estimer les zones responsables des troubles du langage chez des patients ayant eu un AVC, directement à partir des données. L'abord de ce problème se fait dans un cadre bayésien. Afin de tenir compte des relations inhérentes entre voxels voisins, nous utilisons un champ de Markov. Nous comparons la méthode proposée avec celles de la littérature. Les résultats montrent que l'approche proposée est pertinente, surtout quand le nombre de patients est réduit.

Mots-clés. Estimation statistique, Modélisation bayésienne, champs de Markov, VLSM, AVC, test LAST.

Abstract. The study of brain-injured patients is a powerful paradigm for investigating structure–function relationships using neuroimaging techniques. Voxel-based Lesion-Symptom Mapping (VLSM) has been widely used to detect structure–function associations in neuroimaging studies [BWS+03]. This requires using registered images from a lot of patients and performing statistical tests voxel-by-voxel. It is then possible to infer on the areas that are involved on a given disorder (e.g, language disorder). However the VLSM approach is based on Student t-test

for which normality does not always hold and therefore often requires a large number of patients. The aim of this paper is to propose a new approach that allows to reduce the number of patients while maintaining the quality of the results. We have proposed in [FLP18] an alternative nonparametric and Bayesian test using Pólya trees. However, this method, as those in the literature, treats each voxel indepedently of its neighbors. In this paper, we propose a completely different approach. We view the problem as a statistical estimation one, rather than testing. We aim at estimating the language areas from stroke patients, directly from the data. The estimation problem is cast in a Bayesian framework. In order to take into account the spatial correlation between neighboring voxels, the proposed Bayesian model is equipped with a Markov random field. The proposed method is compared to those of the literature. The results highlight that the proposed approach is relevant, especially when the number of patients is reduced.

Keywords. Statistical Estimation, Bayesian Modeling, Markov random field, VLSM, stroke, LAST test.

1 Introduction

Les troubles du langage sont une des sérieuses complications à la suite d'un AVC. D'après les études épidémiologiques, 25-30% des patients ayant fait un AVC développent de tels troubles. Cependant, la détermination des régions impliquées se fait toujours dans le même cadre : un test statistique effectué voxel par voxel et permettant de choisir entre deux hypothèses, l'hypothèse nulle étant que le voxel n'a pas d'effet sur la fonction du langage. Les voxels dits significatifs (pour lesquels H_0 est rejetée) sont considérés comme étant impliqués dans de tels troubles. Dans [FLP18], nous avons proposé comme alternative aux tests fréquentistes communément utilisés, un test statistique bayésien non paramétrique qui a démontré sa pertinence, comparé aux tests fréquentistes, surtout lorsque le nombre de patients est réduit. Cependant, nous ne tenions pas compte de la corrélation spatiale qui peut exister entre voxels voisins. Pour ce faire, on utilise un champ de Markov, plus précisément celui d'Ising, qui introduit une régularisation spatiale, convenable surtout lorsque le nombre de patients est réduit.

2 Données

On considère N patients indicés $n = 1, \dots N$. Pour chaque patient, on dispose d'une image IRM de son cerveau (divisée en petits volumes d'environ 1 mm^3 appelés voxels) et de son score au test du langage LAST (Language Screening Test [FRFR$^+$11]). Pour les images, les lésions sont délinées manuellement ou automatiquement sur chacune d'elles. Cela permet d'avoir une image binaire des lésions des patients. Pour chaque voxel j ($j \in \{1, \dots, J\}$) et chaque patient n, on dispose donc d'un indicateur binaire $z_{n,j}$ tel que :

$z_{n,j} = 1$ si le patient n présente une lésion au voxel j; $\quad z_{n,j} = 0$ sinon.
Puis, un recalage des images est effectué sur un cerveau type afin de faire coïncider les régions. Considérons maintenant les scores au test LAST et notons Y la variable donnant le score. Dans le LAST, si le score est strictement inférieur à 15, le patient est considéré comme aphasique. Soit

$y_n \in \{0, \ldots, 14\}$ si le patient n est aphasique, $\quad y_n = 15$ sinon.

Nos observations sont donc constituées de ces images des lésions et des scores au test LAST. Notre objectif est, à partir de ces observations, d'inférer sur les régions impliquées dans les troubles du langage. Dans la suite, on désignera de telles régions sous le vocable "régions du langage" ou "zones du langage". Nous présentons maintenant le modèle proposé.

3 Modèle d'estimation bayésienne

Avant d'aller plus loin introduisons quelques notations. Soit J_1 le nombre (inconnu) de voxels dans les régions du langage et J_2 son complément ($J_1 + J_2 = J$). Soit N_1 (resp. N_2) le nombre de patients aphasiques (resp. non-aphasiques), avec $N_1 + N_2 = N$; θ_1 et θ_0 désignant respectivement les probabilités de lésions dans les zones du langage chez les aphasiques et les non-aphasiques. Enfin, soit θ la probabilité qu'un voxel en dehors des zones du langage soit lésionnée. On définit une variable binaire non observée indiquant l'implication ou non du voxel. Soit

$$\omega_j = 1 \text{ si le voxel } j \text{ appartient aux zones du langage}\,; \quad \omega_j = 0 \text{ sinon.}$$

Puisque dans le test LAST tous les patients ayant un score strictement inférieur à 15 sont considérés comme aphasiques, on propose de seuiller les scores comme suit :

$$\bar{y}_n = 1 \text{ si l'individu } n \text{ est aphasique}\,; \quad \bar{y}_n = 0 \text{ sinon.}$$

Introduisons maintenant la vraisemblance et les lois *a priori* proposées pour les paramètres.

Vraisemblance Les $z_{n,j}$ sont i.i.d. et suivent des lois de Bernoulli,

$$\begin{aligned} z_{n,j}|\omega_j = 0 &\sim \text{Ber}(\theta) \\ z_{n,j}|\omega_j = 1, \bar{y}_n = 1 &\sim \text{Ber}(\theta_1) \\ z_{n,j}|\omega_j = 1, \bar{y}_n = 0 &\sim \text{Ber}(\theta_0) \end{aligned}$$

L'ensemble des paramètres inconnus est $\boldsymbol{\Psi} = \{\boldsymbol{\omega}, \theta, \theta_0, \theta_1\}$ avec $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_J]^T$.
On introduit maintenant les lois *a priori* proposées pour ces paramètres.

Modélisation *a priori* On suppose que les paramètres $\boldsymbol{\omega}, \theta, \theta_0$ et θ_1 sont *a priori* indépendants entre eux et on considère des lois Beta pour les probabilités de lésions :

$$\theta \sim \text{Beta}(a, b); \quad \theta_0 \sim \text{Beta}(\alpha_0, \beta_0); \quad \theta_1 \sim \text{Beta}(\alpha_1, \beta_1).$$

Pour le paramètre d'intérêt $\boldsymbol{\omega}$ (qui est plutôt une variable latente), il est naturel de considérer qu'il y a une certaine corrélation entre les probabilités $\mathrm{P}\left[\omega_j = k\right]$, pour le voxel j et ceux de ses voisins. Pour ce faire, on considère le modèle d'Ising, de paramètre de champ β, donné par :

$$\mathrm{P}\left[\omega_j = k|\boldsymbol{\omega}_{\setminus j}\right] \propto \exp\left[\sum_{j' \in \mathcal{V}(j)} V_2(k, j')\right] \qquad k \in \{0, 1\} \qquad \text{où}$$

- $\boldsymbol{\omega}_{\setminus j}$ désigne le vecteur $\boldsymbol{\omega}$ privé de son j-ème élément,
- $\mathcal{V}(j)$ est l'ensemble des indices des voisins du voxel j,
- $V_2(k, j')$: sert à promouvoir la régularisation spatiale.

$$V_2(k, j') = \beta\delta(k - \omega_{j'}),$$

où δ est le symbole de Kronecker. La quantité de corrélation spatiale du champ d'Ising est définie par le coefficient β. En effet, il détermine les interactions entre voxels voisins. Si $\beta = 0$, les voxels voisins sont indépendants. On considèrera $\beta > 0$. La loi *a priori* correspondante pour $\boldsymbol{\omega}$ peut s'écrire :

$$\mathrm{P}\left[\boldsymbol{\omega}|\beta\right] = Z(\beta)^{-1} \exp\left[\sum_{j' \in \mathcal{V}(j)} \beta\delta(k - \omega_{j'})\right] \qquad k \in \{0, 1\}$$

La constante de normalisation $Z(\beta)$ est problématique dans le modèle d'Ising. Cependant, si β est connu, elle peut être ignorée. Nous travaillons ici avec un β fixé *a priori* .

Inférence : Ayant supposé des lois *a priori* indépendantes entre paramètres et du fait des lois conjuguées choisies, l'inférence sur les paramètres se fait aisément via un échantillonneur de Gibbs. L'algorithme tire successivement suivant les lois conditionnelles suivantes :

- **Loi conditionnelle de θ :** il s'agit d'une loi Beta de paramètres $a + \sum_{n=1}^{N}\sum_{j=1}^{J_2} z_{nj}$ et $b + NJ_2 - \sum_{n=1}^{N}\sum_{j=1}^{J_2} z_{nj}$.
- **Loi conditionnelle de θ_1 :** c'est aussi une Beta de paramètres $\alpha_1 + \sum_{n=1}^{N_1}\sum_{j=1}^{J_1} z_{nj}$ et $\beta_1 + N_1J_1 - \sum_{n=1}^{N_1}\sum_{j=1}^{J_1} z_{nj}$.
- **Loi conditionnelle de θ_0 :** là aussi une Beta de paramètres $\alpha_0 + \sum_{n=1}^{N_2}\sum_{j=1}^{J_1} z_{nj}$ et $\beta_0 + N_2J_1 - \sum_{n=1}^{N_2}\sum_{j=1}^{J_1} z_{nj}$.
- **Loi conditionnelle de $\boldsymbol{\omega}$:** Le vecteur $\boldsymbol{\omega}$ peut être mis à jour coordonnée par coordonnée. Pour chaque voxel $j \in \{1, 2, \ldots, J\}$, ω_j est une variable aléatoire binaire dont la distribution conditionnelle est caractérisée par

  $$\mathrm{P}\left[\omega_j = 1|\beta, \boldsymbol{\omega}_{-j}, \mathbf{z}, \bar{\mathbf{y}}\right] = \frac{\mathrm{P}\left[\omega_j = 1|\beta, \boldsymbol{\omega}_{-j}\right] f(\mathbf{z}|\omega_j = 1, \beta, \bar{\mathbf{y}})}{\mathrm{P}\left[\omega_j = 1|\beta, \boldsymbol{\omega}_{-j}\right] f(\mathbf{z}|\omega_j = 1, \beta, \bar{\mathbf{y}}) + \mathrm{P}\left[\omega_j = 0|\beta, \boldsymbol{\omega}_{-j}\right] f(\mathbf{z}|\omega_j = 0, \beta, \bar{\mathbf{y}}))}$$

  avec $\mathrm{P}\left[\omega_j = k|\beta, \boldsymbol{\omega}_{-j}\right] \propto \exp\left(\beta \sum_{j' \in \mathcal{V}(j)} \delta(k - \omega_{j'})\right)$, $k \in \{0, 1\}$;

$$f(\mathbf{z}|\omega_j = 0, \beta, \bar{\mathbf{y}}) = \prod_{n=1}^{N} \theta^{z_{nj}} (1-\theta)^{1-z_{nj}} = \theta^{\sum_{n=1}^{N} z_{nj}} (1-\theta)^{N-\sum_{n=1}^{N} z_{nj}};$$
$$f(\mathbf{z}|\omega_j = 1, \beta, \bar{\mathbf{y}}) = \prod_{\{n:\bar{y}_n=1\}} \theta_1^{z_{nj}} (1-\theta_1)^{1-z_{nj}} + \prod_{\{n:\bar{y}_n=0\}} \theta_0^{z_{nj}} (1-\theta_0)^{1-z_{nj}}.$$

4 Résultats

On a analysé les données de 58 patients (47 hommes et 11 femmes) présentant une lésion dans l'hémisphère gauche du cerveau. Tous ces patients ont eu un AVC ischémique en phase aiguë (< 7 jours). La moyenne d'âge de ce groupe est de 66.1 ans (SD=13.4, range=19-91). Le recrutement des patients a été fait au service de neurologie du centre hospitalier régional d'Orléans. Ces patients ont passé le test LAST et on dispose aussi des images binaires de leurs lésions. L'objectif est d'évaluer la performance de la méthode d'estimation proposée et de le comparer aux tests fréquentistes communément utilisés : le test de Student (utilisé dans le VLSM de [BWS$^+$03]), et deux tests non paramétriques, à savoir les tests de Kolmogorov-Smirnov et de Mann-Whitney. Nous le comparons aussi avec le test bayésien non paramétrique que nous avions proposé dans nos travaux initiaux, et désigné BNP-PT. Les images sont traitées coupe par coupe et correspondent donc à des images 2D de taille 181×227.

Pour la méthode bayésienne proposée, on utilise un voisinage de 4 pixels. Le paramètre du champ d'Ising est $\beta = 2.2$. Les hyperpamètres ont été choisis de la façon suivante : $a = b = \alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 0.001$. Ces choix ont été faits en fonction des résultats obtenus sur données simulées et non présentés ici.

Nous avons considéré 1000 itérations de l'échantillonneur de Gibbs et écarté les 500 premières pour la période de burn-in. Il est à noter que l'algorithme converge rapidement (au bout de 100 itérations). En considérant la moyenne sur les itérations post burn-in, on obtient les estimées suivantes pour les paramètres : $\theta = 0.0072$, $\theta_0 = 0.0012$ et $\theta_1 = 0.1543$. En ce qui concerne l'estimée de l'image $\boldsymbol{\omega}$, on a choisi la MPM (Marginal Posterior Mode), au lieu de celle de celle communément utilisée qu'est la moyenne *a posteriori* . En effet, pour plusieurs types d'images, particulièrement les binaires, l'utilisation de la moyenne *a posteriori* ne fournit pas de résultats adéquats. Une alternative est donc l'estimée image MAP (Maximum A Posteriori) ou la MPM (Marginal Posterior Mode). On utilise cette dernière. L'estimée $\hat{\boldsymbol{\omega}}_{MPM}$ peut être obtenue en utilisant N échantillons MCMC post burn-in et calculer

$$\hat{\boldsymbol{\omega}}_{MPM,i} = \begin{cases} 1 \text{ si le nombre de fois où } \omega_i \text{ est égal à 1 supérieur ou égal à } N/2, \\ 0 \text{ sinon.} \end{cases}$$

Nous avions d'abord regardé les résultats obtenus avec $n = 58$ patients. Les résultats (non reproduits ici du fait de l'espace limité) ont montré que les méthodes mises en compétition localisaient plus ou moins les aires classiques du langage, à savoir les aires dites de Broca et

de Wernicke. Classiquement, l'aire de Broca est la zone impliquée dans la production des mots parlés tandis que l'aire de Wernicke est associée à la compréhension de ces mots. Puisque notre objectif était de réduire le nombre de patients inclus dans l'étude, nous avons réduit la taille de l'échantillon augmentant ainsi l'incertitude. Les résultats obtenus avec $n = 34$ sont montrés à la figure 1. Les lignes de haut en bas montrent les résultats obtenus avec le VLSM classique (test de Student) ; test de Mann-Whitney ; test de Kolmogorov-Smirnov ; test BNP-PT et la nouvelle méthode d'estimation proposée. Les colonnes correspondent à différentes vues du cerveau. Les aires du langage sont localisées dans l'hémisphère gauche du cerveau (à droite sur les images de la figure 1), de la partie antérieure (Broca) à la postérieure (Wernicke). Les tests fréquentistes ne retrouvent pas une partie de l'aire de Wernicke (test de Student et Mann-Whitney), voire une partie des deux aires (Kolmogorov). Seul le test BNP-PT produit des régions stables lorsque l'on réduit la taille de l'échantillon. La nouvelle méthode d'estimation proposée localise non seulement les aires de Broca et de Wernicke, mais aussi le *gyrus supramarginal*, une troisième région moins connue mais indispensable au langage. Ce résultat est satisfaisant et prometteur, surtout au vu de la petitesse de l'échantillon. Pour finir, nous soulignons tout de même que les résultats dépendent du paramètre du champ d'Ising (fixée ici *a priori* , se basant sur les données simulées). Une perspective de travail est l'estimation de ce paramètre et l'extension des travaux au 3D.

Références

[BWS+03] E. Bates, S. M Wilson, A. Saygin, F. Dick, M. Sereno, R. T Knight, and N. Dronkers. Voxel-based lesion-symptom mapping. *Nature Neuroscience*, 6(5) :448–450, 2003.

[FLP18] M. D. Fall, É. Lavau, and P.Auzou. Voxel-Based Lesion-Symptom Mapping : : A nonparametric Bayesian approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1050–1054, 2018.

[FRFR+11] C. Flamand-Roze, B. Felissard, E. Roze, L. Maintigneux, J. Beziz, A. Chacon, C. Join-Lambert, D. Adams, and C. Denier. Validation of a new language screening tool for patients with acute stroke : the Language Screening Test (LAST). *Stroke*, pages 1224–1229, 2011.

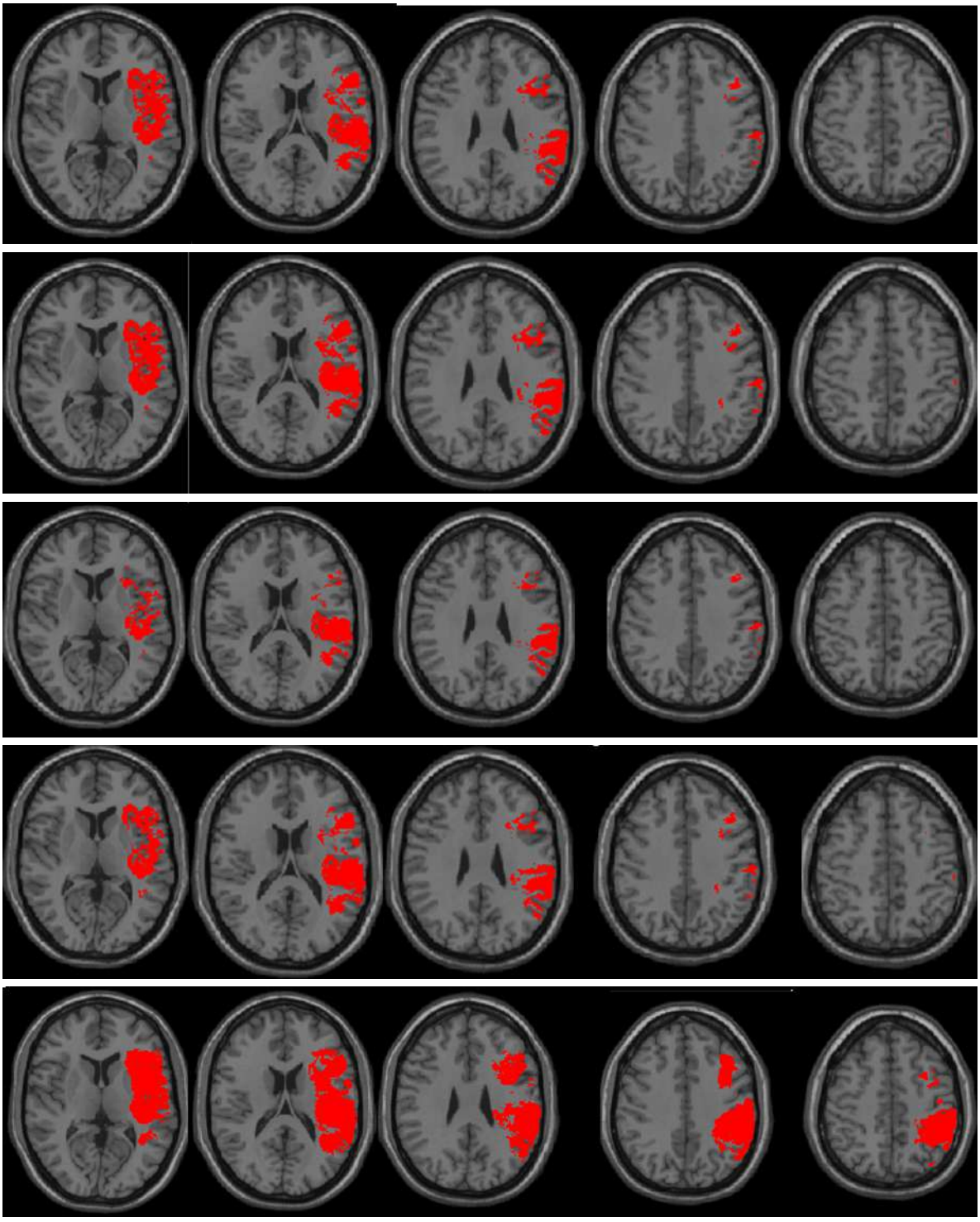


FIGURE 1 – Résultats obtenus par les différentes méthodes mises en compétition.

Experimental Comparison of Semi-parametric, Parametric, and Machine Learning Methods for Time-to-Event Analysis Through the IPEC Score

Camila Fernández[1,2], Chung Shue Chen[2], Pierre Gaillard[3], Alonso Silva[4]

[1]*Sorbonne Université,* [2]*Nokia Bell Labs,* [3]*INRIA,* [4]*Safran Tech*
{*camila.fernandez, chung_shue.chen*}*@nokia.com, pierre.gaillard@inria.fr, alonso.silva-allende@safrangroup.com*

Abstract. In this paper, we make an experimental comparison of semi-parametric (Cox proportional hazards model, Aalen additive model), parametric (Weibull AFT model), and machine learning methods (Random Survival Forest, Gradient Boosting Cox proportional hazards loss, DeepSurv) through the IPEC score on three different datasets (PBC, GBCSG2 and TLCM).

Keywords. Machine Learning, Survival Analysis, Health, Bootstrap, IPEC score.

Acknowledgment. The work presented here has been partially carried out at LINCS.

1 Introduction

Time-to-event analysis is a branch of statistics that looks for modeling the time remaining until a certain critical event occurs. For example, this event can be the time until a biological organism dies or the time until a machine fails. There are many other examples, in healthcare, the aim is usually to predict the time until a patient with certain disease dies or the time until the recurrence of an illness, whereas in telecom, the goal could be to predict the customer churn, etc. One of the main interests of time-to-event analysis is right censoring, it comes naturally from the fact that not necessary all the samples have reached the event time which makes the problem more difficult and a different challenge from the typical regression problem.

In Fernandez et all. (2020), the performance (through the concordance index) of several models have been compared on two different datasets, both of them related to a healthcare approach. The first one is about patients diagnosed with primary biliary cirrhosis (PBC) where the goal is to predict the time until the patient dies. The second dataset consists on patient diagnosed with breast cancer and the objective is to predict the recurrence of the disease. Here we add a third dataset which is from a different source, it consists on clients from a telecommunication company, Telco (TLCM), and the aim is to predict the customer churn. We also consider a different score to carry out this comparison, the IPEC score (see Section 1.2).

In this work, we implement a bootstrapping technique for the computation of the IPEC score in the test set, this is with the aim of obtain a better approximation of the asymptotic behavior of the estimated event time. Each sample of each dataset has an

observed time which can correspond either to a survival time or a censored time. A censored time will be a lower bound for the survival time and so we will be in the case in which the critical event has not occurred at the moment of the observation.

Survival and hazard function The fundamental task of time-to-event analysis is to estimate the probability distribution of time until some event of interest happens.

Consider a covariates/features vector X, a random variable that takes on values in the covariates/features space $\mathcal{X}$. Consider a survival time T, a non-negative real-valued random variable. Then, for a feature vector $x \in \mathcal{X}$, our aim is to estimate the conditional survival function:

$$S(t|x) := \mathbb{P}(T > t|X = x), \tag{1}$$

where $t \geq 0$ is the time and $\mathbb{P}$ is the probability function. In order to estimate the conditional survival function $S(\cdot|x)$, we assume that we have access to a certain dataset in which for the i-th sample we have: X_i the feature vector, δ_i the survival time indicator, which indicates whether we observe the survival time or the censoring time, and Y_i which is the survival time if $\delta_i = 1$ and the censoring time otherwise. We split the dataset into a training set of size n and a test set of size m. The training set is used to estimate the parameters of each model and the test set to measure how accurate is the estimation of the probability function.

Many models have been proposed to estimate the conditional survival function $S(\cdot|x)$ such as Cox proportional hazards from Cox (1972), gradient boosting from Friedman (2001) and random survival forest from Ishwaran (2008). The most standard approaches are the semi-parametric and parametric models, which assume a given structure of the hazard function $h(t|x) := -\frac{\partial}{\partial t} \log S(t|x)$.

IPEC score The IPEC score, introduced first by Gerds and Schummacher (2006), is an alternative score to the concordance index that we used in a previous work in Fernandez et all. (2020) in order to measure the accuracy of time to event models. The IPEC score is a consistent estimator for the mean square error of the probability function S. We used a variant of the original IPEC score which was presented by Chen (2019). This score approximates the following MSE of a survival probability estimator $\hat{S}$, which cannot be directly computed from the dataset,

$$MSE(\hat{S}) = \int_0^\tau \mathbb{E}[(\mathbb{1}\{T > t\} - \hat{S}(t|X))^2]dt \tag{2}$$

where τ is a user-specified time horizon and T is the survival time of feature vector X. Let us define $S_C(t|x) = \mathbb{P}(C > t|X = x)$ where C is the censored time. Then, the IPEC score is computed as follows. Let $\hat{S_C}$ be an estimator of S_C,

$$IPEC(\hat{S}) = \frac{1}{m} \sum_{i=1}^{m} \int_0^\tau W_i(t)(\mathbb{1}\{Y_i > t) - \hat{S}(t|X_i))^2 \tag{3}$$

where (X_i, Y_i, δ_i) for $0 < i \leq m$ are the samples of the test set. W_i is defined as:

$$W_i(t) = \begin{cases} \frac{\delta_i \mathbb{1}\{Y_i \leq t\}}{\hat{S}_C(Y_i|X_i)} + \frac{\mathbb{1}\{Y_i > t\}}{\hat{S}_C(t|X_i)} & \hat{S}_C(t|X_i) \geq \theta \\ 1/\theta & \text{otherwise.} \end{cases} \tag{4}$$

Here, θ is an user-specified bound which was introduced in order to prevent a division by 0 and then, in the worst case, the IPEC score is finite. The addition of this last parameter is the only difference between the original IPEC score of Gerds and Schumacher (2006) and the variant introduced by Chen (2019). In practice, θ can be set as an arbitrarily small but positive constant. Note that $0 \leq IPEC(\hat{S}) \leq \tau/\theta$.

2 Datasets Description

German Breast Cancer Study Group dataset (GBCSG2) The German Breast Cancer Study Group (GBCSG2) dataset, made available by Schumacher et al. (1994), studies the effects of hormone treatment on recurrence-free survival time. The event of interest is the recurrence of cancer time. The dataset has 686 samples and 8 covariates/features: age, estrogen receptor, hormonal therapy, menopausal status (premenopausal or postmenopausal), number of positive nodes, progesterone receptor, tumor grade, and tumor size. At the end of the study, there were 387 patients (56.4%) who were right censored (recurrence-free).

Mayo Clinic Primary Biliary Cirrhosis dataset (PBC) The Mayo Clinic Primary Biliary Cirrhosis dataset, made available by Therneau and Grambsch (2000), studies the effects of the drug D-penicillamine on the survival time. The event of interest is the death time. The dataset has 276 samples and 17 covariates/features: age, serum albumin, alkaline phosphatase, presence of ascites, aspartate aminotransferase, serum bilirunbin, serum cholesterol, urine copper, edema, presence of hepatomegaly or enlarged liver, case number, platelet count, standardized blood clotting time, sex, blood vessel malformations in the skin, histologic stage of disease, treatment and triglycerides. At the end of the study, there were 165 patients (59.8%) who were right censored (alive).

Kaggle Telco Churn (TLCM) The Kaggle Telco Churn dataset, made available by Kaggle in 2008, studies the possible causes of customer churn in a telecommunication enterprise. The event of interest is the churn time of the clients. The dataset has 7043 samples and 19 covariates/features: customer ID, gender, senior citizen, partner, dependents, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges and total charges. At the end of the study, there were 5174 clients (73%) who were right censored (not have churned yet).

3 Models

We took into consideration several models for the comparison analysis. Semi-parametric models such as Cox (1972) and Aalen's additive (1989), both models assume certain parametrical structure on the hazard function. These models are semi-parametric in the sense that the baseline hazard function does not have to be specified and it can vary allowing a different parameter to be used for each unique survival time.

We also consider a parametric model, Weibull accelerated failure time by Liu (2018), it supposes that the hazard function depends on an accelerated rate $\lambda(x)$ which can be estimated parametrically.

Finally we consider machine learning models such as Random survival forest proposed by Ishwaran et all. (2008), Gradient boosting cox proportional hazards loss proposed by Friedman (2001), DeepSurv by Katzman et all. (2018) and a variation of random survival forest proposed by Chen (2019). We also considered a randomized search of the parameters which was done by cross validation. For more information and details about these models look at Fernandez et all. (2020).

4 Results and Conclusion

For each dataset, we choose 25 different seeds for splitting the data which generates 25 different partitions between training and test sets (75% and 25% respectively). We repeat the experiment 25 times and we make a boxplot with the distribution of the IPEC scores obtained. Fig. 1, 3 and 5 respectively compare the IPEC score for PBC, GBCSG2 and TLCM datasets, and Fig. 2, 4 and 6 show the same comparison after re-sampling the test set five times (bootstrapping).

In Fig. 1, we can appreciate that Gradient boosting with randomized search of the parameters performs better than the other models and DeepSurv is in second place. Fig. 3 shows that DeepSurv outperforms all the other models for GBCSG2. And finally, Fig. 5 shows the comparison for TLCM dataset where we can observe that Cox proportional Hazards model is the model with the best performance and DeepSurv dropped down to the fifth place.

Furthermore, we can observe that traditional methods performed reasonably well for the big dataset TLCM, but they underperformed against machine learning methods for the smaller datasets (GBCSG2 and PBC). We can also observe that the deep learning method (Deepsurv) performed better than random survival forest model in all the datasets.

Fig. 2, 4 and 6 show the comparison of the IPEC score using the bootstrapping technique. We appreciate that DeepSurv outperforms all the other models for the smaller datasets (PBC and GBCSG2) and Cox proportional hazards has the best result for the biggest dataset (TLCM) as in the previous case without re-sampling.

This shows that there is no much difference in the results when we apply the bootstrapping technique for the test of the models. In addition, we know that classical methods

are easier to interpret in the sense of measure how each covariate/feature influences in the model. For the case of PBC dataset, gradient boosting with random search outperforms DeepSurv by a 12% while similarly in TLCM the method cox proportional hazards increase the performance by a 12% with respect to Deepsurv model. The case of GBCSG2 is different because DeepSurv improves the performance in a 37% compared to Aalen's additive method, therefore, if this increment of performance is significant enough to compensate the loss of interpretation will depend mainly on the applications.

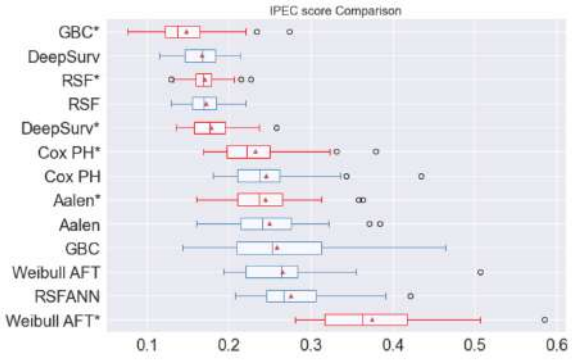


Figure 1: IPEC score comparison for PBC dataset

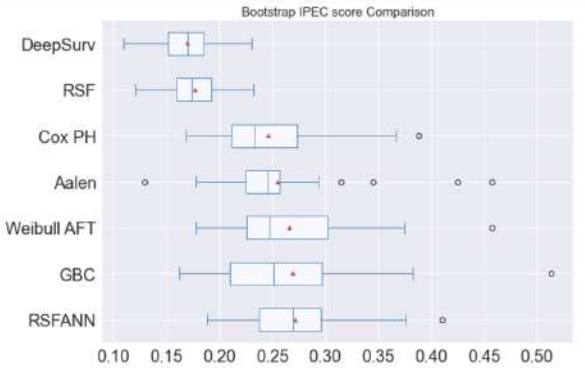


Figure 2: IPEC score comparison with bootstrapping

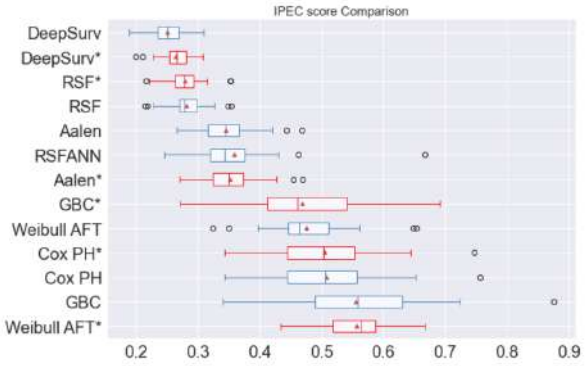


Figure 3: IPEC score comparison for GBCSG2 dataset

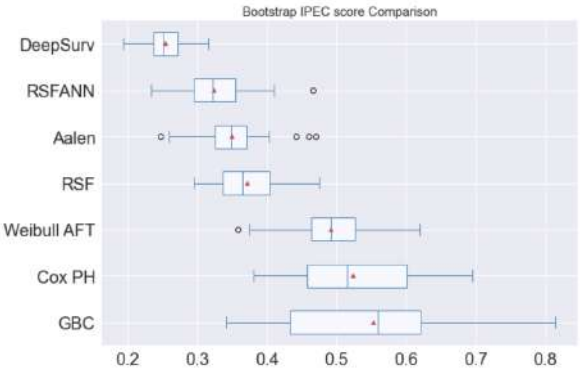


Figure 4: IPEC score comparison with bootstrapping

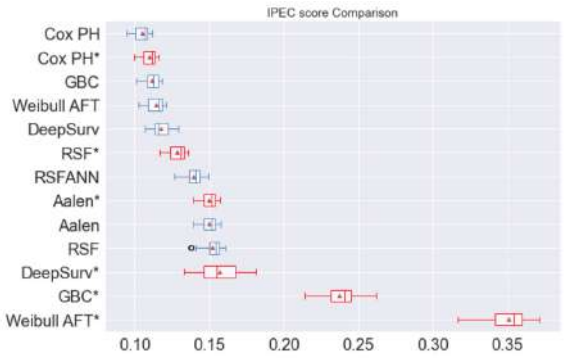


Figure 5: IPEC score comparison for TLCM dataset

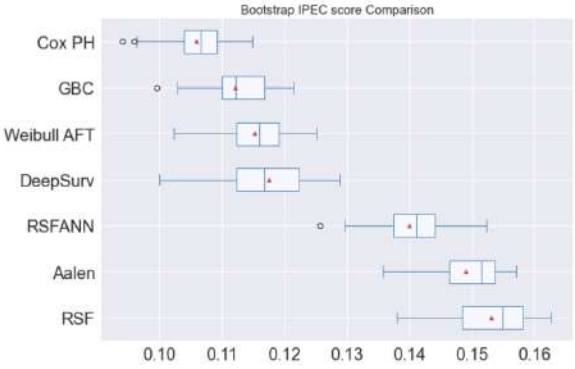


Figure 6: IPEC score comparison with bootstrapping

Bibliographie

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, vol. 8, pp. 907–925.

Chen, G. (2019). Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates. arXiv preprint arXiv:1905.05285.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, Series B. 34 (2): 187–220.

Davidson-Pilon, C., et al. (2020), CamDavidsonPilon/lifelines:v0.23.9, doi: 10.5281/zenodo.805993.

Efron, B. (1982), The jacknife, the bootstrap and other resampling plans, *SIAM*.

Fernández, C., Chen, C. S., Gaillard, P., Silva, A. (2020), Experimental comparison of semi-parametric, parametric and machine learning models for time-to-event analysis through the concordance index. *Journée de Statistique SFdS,* pp.317-325

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* pp.1189-1232

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860.

Katzman J., et al. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology.*

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python, *JMLR.*

Pölsterl, S. (2019). Scikit-survival:v0.11, doi:10.5281/zenodo.3352342.

Schumacher, M., et al. (1994), Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10), pp. 2086–2093.

Liu, E. (2018), Using Weibull accelerated failure time regression model to predict survival time and life expectancy. *BioRxiv.*

Spatial segmentation of count data with a Bayesian nonparametric Hidden Markov model: application to traffic crash risk mapping

J.-B. Durand[1] & F. Forbes[1] & C.D. Phan[2] & L. Truong[2] & H.D. Nguyen[2] & F. Dama[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France. Florence.Forbes@inria.fr*

[2] *School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia*

Résumé. Nous étudions l'utilisation de modèles bayésiens non paramétriques (BNP) couplés à des champs aléatoires de Markov (MRF) dans un contexte de gestion des risques, pour construire des partitions du risque en régions spatialement homogènes. Contrairement à la plupart des méthodes existantes, l'approche proposée ne nécessite pas un choix arbitraire du nombre de classes de risque et détermine automatiquement leurs niveaux de risque. Nous proposons un modèle appelé BNP Hidden MRF (BNP-HMRF) qui est capable de gérer des données de comptage. L'inférence du modèle est effectuée à l'aide d'un algorithme variationnel Bayes Expectation - Maximization et l'approche est illustrée sur des données d'accidents de la circulation dans l'état de Victoria, en Australie. Les résultats obtenus corroborent bien avec la littérature sur la sécurité routière. Plus généralement, le modèle présenté ici pour la cartographie des risques offre un moyen efficace, pratique et rapide de procéder à la partition de données de comptage spatialement localisées.

Mots-clés. Sécurité routière; Accidents de la circulation; Cartographie du risque; Modèles non paramétrique bayésien; Champ aléatoire de Markov; Algorithme VBEM

Abstract. We investigate the use of Bayesian nonparametric (BNP) models coupled with Markov random fields (MRF) in a risk mapping context, to build partitions of the risk into homogeneous spatial regions. In contrast to most existing methods, the proposed approach does not require an arbitrary commitment to a specified number of risk classes and determines their risk levels automatically. We consider settings in which the relevant information are counts and propose a so called BNP Hidden MRF (BNP-HMRF) model that is able to handle such data. The model inference is carried out using a variational Bayes Expectation–Maximisation algorithm and the approach is illustrated on traffic crash data in the state of Victoria, Australia. The obtained results corroborate well with the traffic safety literature. More generally, the model presented here for risk mapping offers an effective, convenient and fast way to conduct partition of spatially localised count data.

Keywords. Road safety; Traffic crashes; Risk mapping; Bayesian nonparametrics; Markov random field; Variational Bayes Expectation–Maximisation algorithm

1 Introduction

Traffic-related injuries and deaths are major problems in contemporary societies. Social economic losses from traffic crashes, in particular from motor vehicle crashes, are enormous. This makes road and traffic safety a major concern, worldwide. The nondecreasing relationship between crash casualties and population suggests that safety improvements could be gained from a better prediction of crash occurrences. Traffic crashes are complex events involving the interactions of various factors. In particular, since road transport involves distances by nature, most studies call for spatial analysis to account for geographical locations and environments in which crashes occur. The goal is often to accurately predict the risks at different locations [5] and to link these risk values to other variables for interpretability, or to assess the impact of several risk factors [10, 7] and road safety measures [3]. The hope is to identify the potential causal sources of crashes, and to apply appropriate control procedures and protection measures; see e.g., [11]. We primarily aim at highlighting areas with different risk levels, with respect to various covariates, such as population density, traffic density, signalisation density, etc. The interest of such partitioning is to highlight spatial heterogeneity, to locate high risk areas (so called risk hot spots), and to determine whether they exhibit some structure in space that could be analysed or directly interpreted.

In this work, to handle discontinuities in the spatial structure of the risk, without having to arbitrary choose their number, we propose to operate in the framework of Bayesian nonparametric (BNP) methods [4]. More specifically, we build on methods recently proposed for the modelling of continuous observations by [6]. We extend the approach, referred to as BNP-HMRF, to the modelling of count data. We derive the corresponding variational Bayes Expectation–Maximisation (VBEM) algorithm for the model estimation. The approach is then illustrated on traffic crash data in the state of Victoria, Australia. The analysis provides risk zones and risk levels that are globally coherent with other findings in the literature.

2 BNP-HMRF model for count data

The study aims at providing some risk mapping of traffic crashes, based on data regarding geographical zones. Since, on average, the number of traffic crashes increases with respect to other variables characterising the traffic importance (e.g., population size, traffic intensity, length of road network), the numbers of crashes have to be normalised with respect to at least one of these variables. The obtained ratio provide what we interpret as risks. One objective is then to account for some spatial heterogeneity regarding the observed risks. The model described in the following lines aims at clustering regions with close risks to provide a labelled map, where each label is associated with some risk level.

Considering J regions, where we let y_j represent the number of crashes occurring in region $j \in \{1, \dots, J\}$, characterised by a normalisation variable N_j; e.g., the population

size of region j. For the sake of clarity, we first assume that there is a finite set of K risk levels $\mathbf{\Lambda} = \{\lambda_0, \ldots, \lambda_{K-1}\}$, that are ordered so that λ_k is the $(k+1)$th smallest level. Since the risk level associated to region j is not known in advance, a variable $z_j \in \{0, \ldots, K-1\}$ is introduced to indicate the assigned risk level, i.e. $z_j = k$, when region j is at risk level λ_k. When region j is at risk level λ_k, the number of traffic crashes y_j is then assumed to be Poisson distributed with mean $\lambda_k N_j$. That is, y_j conditioned on $z_j = k$ has probability mass function

$$p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j), \tag{1}$$

where $\mathcal{P}(\cdot; \lambda_k N_j)$ denotes the probability mass function of the Poisson distribution with parameter $\lambda_k N_j$. From a generic point of view, $p(y_j|z_j = k; \mathbf{\Lambda}, N_j)$ is referred to as an emission distribution. As a consequence, the mean number of crashes is a linear function of N_j: $\mathrm{E}[y_j|z_j = k; \mathbf{\Lambda}, N_j] = \lambda_k N_j$. The goal is then to estimate the risk levels $\mathbf{\Lambda} = \{\lambda_0, \ldots, \lambda_{K-1}\}$ and the most likely risk mapping through the most likely values of the z_js. In practice, risk levels are likely to vary smoothly across regions. It is more likely that neighbouring regions have the same risk level with possible abrupt changes from a region to another if they have contrasting characteristics. Thus, for a better estimation of risk levels, a Markov random field (MRF) model is used for the set of labels $\boldsymbol{z} = \{z_j, j \in J\}$, to account for spatial dependencies between connected regions. Formally, the regions are seen as the vertices of a graph G. They are connected by an edge in the graph whenever they share a boundary, although other types of connections could be considered (e.g., they either share a boundary or have a common neighbour, etc.). The probability for neighbouring regions having either a similar or different label is controlled by some scalar positive parameter denoted by β. The higher the value of β, the more likely neighbouring regions are at the same risk level.

The number K of risk levels is not usually known in advance and has to be chosen adaptively by users. To avoid this commitment to a fixed number K, we propose an extension of the model that does not restrict the levels to a finite number K. This extension is based on so called Dirichlet Process Mixtures [6] and is referred to as a Bayesian Non-Parametric Hidden Markov Random Field (BNP-HMRF).

The BNP-HMRF model is defined as follows. The set of J regions under consideration is associated to a graph structure $G = (J, E)$, where each $j \in J$ corresponds to a node of G and the set of edges E represents all pairs of regions with a common boundary. The likelihood part of the model is given by (1). The observations are counts $\boldsymbol{y} = \{y_j, j \in J\}$ distributed independently given $\mathbf{\Lambda}$ and $\boldsymbol{z}$ with for every j, $p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j)$. The risk class labels $\boldsymbol{z} = \{z_j, j \in J\}$ are assumed to be distributed as a Markov random field on G with the following distribution:

$$p(\boldsymbol{z}|\beta, \boldsymbol{\pi}) \propto \exp\left(\sum_{j=1}^{J} \ln \pi_{z_j} + \beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i = z_j)}\right) = \left(\prod_{j=1}^{J} \pi_{z_j}\right) \exp\left(\beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i = z_j)}\right), \tag{2}$$

where $\mathbf{1}_{(z_i=z_j)}$ is the indicator function equal to 1 when $z_i = z_j$ and 0 otherwise, $\{i, j\} \in E$ indicates that $\{i, j\}$ is an edge in G, β is some unknown scalar parameter, and the π_ks are weights defined for every $k \geq 0$ as $\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l<k}(1 - \tau_l)$, where $\boldsymbol{\tau}^\top = (\tau_0, \tau_1, \ldots)$ is a sequence of independent, identically distributed (i.i.d.) random variables with distribution Beta$(1, \alpha)$. The parameter α is an hyperparameter which follows a gamma distribution, $\alpha|s_1, s_2 \sim \mathcal{G}(s_1, s_2)$, while each parameter λ_k in (1) is also distributed according to a gamma distribution, $\lambda_k|a_k, b_k \sim \mathcal{G}(a_k, b_k)$.

The construction of the π_ks corresponds to a stick-breaking construction (see Lemma 3.4 in [4] and [9], for details) and guarantees that $\sum_{k=0}^{\infty} \pi_k = 1$, which in turns ensures that the distribution in (2) is a valid Markov field [6]. The complete hierarchical model can thus be stated, for $\boldsymbol{z} = \{z_1, \ldots, z_J\}$ and $k = 0, 1, \ldots$, as follows:

$$
\begin{aligned}
&\lambda_k|a_k, b_k \sim \mathcal{G}(a_k, b_k) \quad \text{(independent)} \\
&\alpha|s_1, s_2 \sim \mathcal{G}(s_1, s_2), \\
&\tau_k|\alpha \sim \mathcal{B}(1, \alpha) \quad \text{(i.i.d.)}, \\
&\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}(1 - \tau_l), \\
&p(\boldsymbol{z}|\boldsymbol{\tau}, \beta) \propto \prod_{j \in J} \pi_{z_j}(\boldsymbol{\tau}) \ \exp(\beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i=z_j)}), \\
&y_j|z_j; \boldsymbol{\Lambda}, N_j \sim \mathcal{P}(y_j; \lambda_{z_j} N_j), \quad \text{for each } j \in J.
\end{aligned}
$$

The model parameters can be estimated using a variational Bayes Expectation–Maximisation algorithm. Details can de found in [1].

3 Application to traffic crash risk mapping

The results presented here comprise of data from Victoria, Australia. Crash data between 2014 and 2018 were obtained from Victoria's open data directory [11].

Since an absolute validation of the risk mapping is difficult, we resort to covariates to assess whether risk level classes encode relevant and contrasted characteristics between different classes. In practice, the set of covariates is the same as the set of possible normalising variables. Four such variables are considered: population size, region size, absolute traffic and traffic density. In this paper, we illustrate our result using the population size. We now consider the population size of a region as N_j, i.e., risks are clustered according to the impact of population size on the number of crashes.

The model led to seven clusters when starting from $K = 10$, among which four have negligible frequencies (see Figure 1). The frequency corresponds to the size of the cluster divided by the total number of regions. Here we do not provide any detailed interpretation for 8 regions in four clusters (clusters labelled as 5 to 9) that essentially are outliers (the

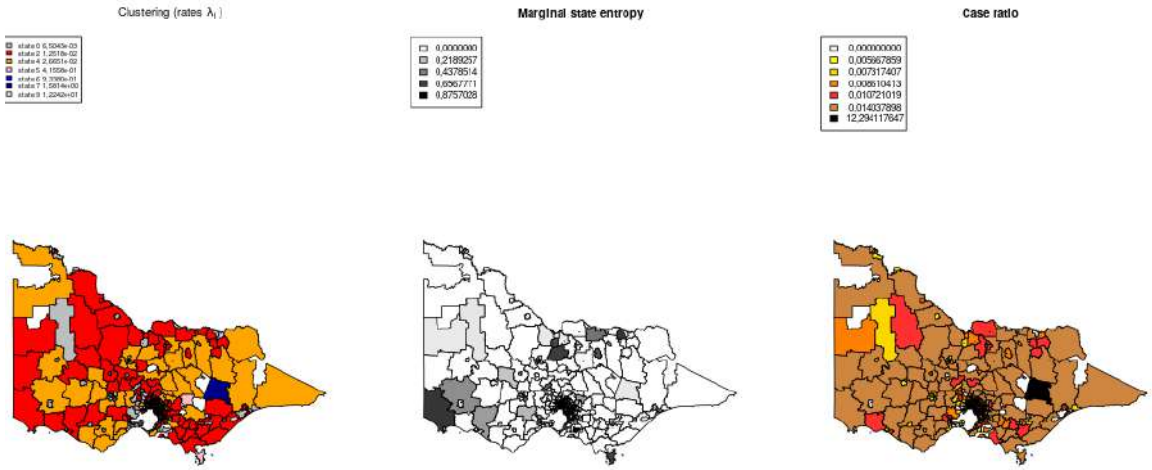


Figure 1: Risk mapping with respect to population size (variable pop16). Right-hand part: segmentation using quantiles on ratio.

regions correspond to zones with 6 to 184 inhabitants, where 9 to 209 crashes occur). The estimate $\hat{\beta} \approx 0.34$ indicates rather low spatial aggregation of clusters.

Risk level 0 is related to peripheral regions that are close to the capital of Melbourne, and enclaves, which are often regional towns or rural centres with substantial residential developments (see Figure 1). These are small regions with high population sizes, high absolute traffic and high traffic densities. Risk level 2 is related to peripheral and central regions. These are medium-sized zones with medium population sizes, absolute traffic and traffic densities. Risk level 4 is related to far peripheral and hypercentral regions (relative to the capital). These are sparsely populated, have varying sizes, with high absolute traffic and traffic densities. The four variables considered are well discriminated by the risk levels, with ANOVA p-values between 10^{-10} and 10^{-15} regarding the effects of the classes (ANOVA not shown).

4 Conclusion

The BNP-HMRF model presented here for risk mapping offers a convenient and fast approach for conducting segmentation of count data regressions indexed by graphs. The proposed model was effective in identifying clusters with distinct risk levels. Detailed analyses of these clusters showed that regions with higher traffic densities tend to have lower traffic density-based crash risk levels, while regions with higher population sizes tend to have lower population-based crash risk levels. These findings corroborate well with the traffic safety literature. It is well-established that crash risks tend to decrease with increasing exposure, such as population or the number of road users [2]. Regarding

further application to traffic crashes, the model could be extended to consider multiple crash exposure variables. However, the possibility to use several risk-normalising variables leads to multiplying classes and makes their interpretation more difficult.

BNP-HMRF models could also be extended to handle multivariate count data, particularly to address modelling problems in ecology where counts correspond to the number of observed species. The emission densities could thus be replaced by Join Species Distribution Models (JSDMs), in which spatial dependencies and heterogeneity sources are usually modelled with univariate CARs [8] but are ignored in the multivariate count data.

References

[1] J.-B. Durand, F. Forbes, C. D. Phan, L. Truong, H. D. Nguyen, and F. Dama. Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia. Feb. 2021.

[2] R. Elvik. *Towards a general theory of the relationship between exposure and risk*. Institute of Transport Economics, Oslo, Norway, 2014.

[3] R. Elvik, T. Vaa, A. Hoye, and M. Sorensen. *The handbook of road safety measures*. Emerald Group Publishing, 2009.

[4] S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[5] D. Lord and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010.

[6] H. Lu, J. Arbel, and F. Forbes. Bayesian nonparametric priors for hidden Markov random fields. *Statistics and Computing*, 2020.

[7] E. Papadimitriou, A. Filtness, A. Theofilatos, A. Ziakopoulos, C. Quigley, and G. Yannis. Review and ranking of crash risk factors related to the road infrastructure. *Accident Analysis & Prevention*, 125:85 – 97, 2019.

[8] Y. Saas and F. Gosselin. Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. *Ecography*, 37(5):476–489, 2014.

[9] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[10] A. Theofilatos and G. Yannis. A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72:244 – 256, 2014.

[11] L. Truong and G. Currie. Macroscopic road safety impacts of public transport: A case study of Melbourne, Australia. *Accident Analysis & Prevention*, 132, 2019.

Titre : La pseudonymisation à la Cour de cassation : l'histoire, les concepts et les évolutions

Auteur : Amaury Fouret
Data scientist, Cour de cassation

Résumé :
La justice doit être accessible et la Cour de cassation s'engage à relever le défi en utilisant les potentialités des technologies appliquées au droit " - Chantal Arens, première présidente de la Cour de cassation.
D'une part, la loi Open Data en France oblige à diffuser les décisions de justice au public. D'autre part, le RGPD impose la protection des données personnelles des citoyens. La Cour de cassation est chargée de la collecte automatisée de la jurisprudence de toutes les juridictions de son périmètre juridictionnel, de sa pseudonymisation et de sa publication.
Après des pilotes avec des acteurs externes, la Cour de cassation a participé au programme « Entrepreneurs d'Intérêt Général » en 2019 et 2020 et a recruté des data scientists, des développeurs et un designer pour construire un module de pseudonymisation basé sur l'apprentissage automatique et une nouvelle interface d'annotation.
Dans cet exposé, nous présenterons le contexte général de notre travail, les choix techniques et les défis. Nous mentionnerons notre boîte à outils et les modules complémentaires personnalisés qui ont permis à notre modèle d’apprentissage profond de bien s’insérer en production. Nous parlerons également de notre méthode de surveillance et de tests, ainsi que de la prédiction de l'incertitude, que nous développons pour assurer la robustesse du modèle. Tous ces changements seront effectués en synergie avec la nouvelle interface d'annotation, qui poussera beaucoup plus loin les possibilités d'améliorer et d'évaluer le modèle d’apprentissage automatique.

RECOMMANDATION ÉQUITABLE VIA UNE PARITÉ STATISTIQUE DANS UN CO-CLUSTERING ORDINAL

Gabriel Frisch, Jean-Benoist Leger & Yves Grandvalet

Université de Technologie de Compiègne, CNRS, Heudiasyc UMR 7253, Compiègne France. {gabriel.frisch,jbleger,yves.grandvalet}@hds.utc.fr

Résumé. Nous présentons un modèle statistique basé sur le LBM ordinal pour réaliser une recommandation sociale. En utilisant une covariable encodant un caractère protégé lié à l'utilisateur, nous obtenons ainsi une mixité proportionnelle au sein des groupes donnant lieu à une recommandation équitable.

Abstract. We present a statistical model based on the ordinal LBM for collaborative filtering. Using a covariate encoding a user protected attribute, we get a statistical parity within groups thus producing a fair recommendation.

Keywords. Statistical parity, fairness, collaborative filtering, latent block model

1 Introduction

Un algorithme est dit équitable entre individus sur la base d'attributs protégés (tel que le sexe, l'âge, ou la catégorie socioprofessionnelle) si ces attributs n'influencent pas le classement, le tri ou la sélection des informations. Dans le cas des systèmes de recommandation, la majorité des modèles utilisés ne permettent pas d'assurer cette équité vis à vis des intérêts des personnes.

Le filtrage collaboratif vise à construire des systèmes de recommandation utilisant l'historique des opinions des utilisateurs. Nous proposons un modèle de filtrage collaboratif basé sur une classification croisée ordinale utilisant une covariable encodant un caractère protégé. En utilisant ce modèle, nous proposons un prédicteur permettant de construire une liste d'objets à recommander indépendante du caractère protégé de l'utilisateur.

Dans cet article, nous présentons notre modèle dans un contexte de filtrage collaboratif. Cependant, il peut être appliqué à tout autre problème de classification croisée avec données manquant au hasard (modèle MAR).

Préliminaires : soient n_1 le nombre d'utilisateurs, n_2 le nombre d'objets de notre système et $\boldsymbol{R} \in \mathbb{R}^{n_1 \times n_2}$ la matrice des évaluations « tacites », supposées continues, où chaque entrée R_{ij} représente la valeur effective ou putative de l'objet j pour l'utilisateur i. Nous utilisons systématiquement l'index i pour les utilisateurs, en ligne sur la matrice $\boldsymbol{R}$, et

l'index j pour les objets, en colonne de la matrice $\boldsymbol{R}$. Lorsque les limites de ces indices ne sont pas détaillées dans les sommes ou produits, i va de 1 à n_1 et j va de 1 à n_2.

Les notes observées $\boldsymbol{R}^{(\mathrm{o})}$ sont supposées être issues d'une quantification des évaluations tacites $\boldsymbol{R}$. Les utilisateurs ne notant pas tous les objets, certaines entrées de $\boldsymbol{R}^{(\mathrm{o})}$ sont manquantes. Soit $\boldsymbol{M} \in \{0,1\}^{n_1 \times n_2}$ la matrice de masque dont chaque entrée M_{ij} indique si la note est observée : $M_{ij} = 1$ si $R_{ij}^{(\mathrm{o})}$ est observée et 0 sinon. Dans un cadre de recommandation, l'objectif consiste à prédire les notes non-observées. La matrice regroupant ces notes manquantes sera notée $\boldsymbol{R}^{(\mathrm{m})}$.

2 Modèle joint des données et du manquement

2.1 Modèle des données complètes

Le modèle à blocs latents (LBM), développé par Govaert et Nadif [1], est un modèle probabiliste génératif permettant de classifier simultanément les lignes et les colonnes d'une matrice de données. Le LBM pose que la matrice de données est structurée en blocs homogènes. Cette structure est rendue visible en réorganisant les lignes et les colonnes selon leurs classes d'appartenance respectives ; pour k_1 classes en ligne et k_2 classes en colonne, le réarrangement révèle $k_1 \times k_2$ blocs homogènes dans la matrice de données.

Notons $\boldsymbol{U}$ la matrice indicatrice $n_1 \times k_1$ d'appartenance des utilisateurs à leur classe en ligne, avec $U_{iq} = 1$ si l'utilisateur i appartient au groupe q. De façon similaire, $\boldsymbol{V}$ désigne la matrice indicatrice $n_2 \times k_2$ d'appartenance des objets aux classes en colonnes. Les appartenances aux classes forment une partition : chaque utilisateur appartient à exactement une classe.

Le modèle que nous proposons est une variante du LBM Gaussien et repose sur plusieurs hypothèses de distribution et d'indépendance :

— Les appartenances des utilisateurs et des objets aux classes sont *a priori* mutuellement indépendantes et identiquement distribuées respectivement selon les multinomiales $\mathcal{M}(1;\boldsymbol{\alpha})$ et $\mathcal{M}(1;\boldsymbol{\beta})$, où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_1})$ et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_2})$ sont les proportions de mélange des groupes en ligne et en colonne :

$$p(\boldsymbol{U}, \boldsymbol{V}) = p(\boldsymbol{U})\; p(\boldsymbol{V}) = \prod_i p(\boldsymbol{U}_i;\boldsymbol{\alpha}) \prod_j p(\boldsymbol{V}_j;\boldsymbol{\beta}) \;\; .$$

— Les utilisateurs et les objets sont responsables d'effets individuels indépendants notés respectivement A_i et B_j, identiquement distribués :

$$A_i \overset{\text{iid}}{\sim} \mathcal{N}\big(0, \sigma_A^2\big), \qquad \sigma_A^2 \in \mathbb{R}_+^* \;\; , \qquad\qquad B_j \overset{\text{iid}}{\sim} \mathcal{N}\big(0, \sigma_B^2\big), \qquad \sigma_B^2 \in \mathbb{R}_+^* \;\; .$$

— Connaissant les classes des utilisateurs et des objets $(\boldsymbol{U}, \boldsymbol{V})$ et les effets individuels

$(\boldsymbol{A}, \boldsymbol{B})$, les notes R_{ij} sont indépendantes et distribuées selon :

$$p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}; \boldsymbol{\pi}) = \prod_{ij} p(R_{ij}|\boldsymbol{U}_i, \boldsymbol{V}_j, A_i, B_j; \boldsymbol{\pi}) \ , \ \boldsymbol{\pi} \in \mathbb{R}^{k_1 \times k_2}$$
$$(R_{ij}|U_{iq}V_{jl} = 1, A_i, B_j) \overset{\text{iid}}{\sim} \mathcal{N}\Big(\pi_{ql} + A_i + B_j, \sigma^2\Big) \ , \ \sigma^2 \in \mathbb{R}_+^* \ . \tag{1}$$

Introduction d'une covariable : l'introduction d'une covariable pour la loi d'émission de $\boldsymbol{R}$ conditionnellement aux variables latentes permet de séparer l'effet de la covariable de l'effet des groupes du coclustering. En introduisant une covariable, notée Cov_i, encodant le sexe de l'individu, nous espérons ainsi obtenir une mixité proportionnelle au sein des groupes ; c'est à dire que la probabilité qu'un individu tiré au hasard soit un homme est la même quelque soit le groupe considéré. Dans cet article, la covariable est liée au sexe de l'individu, mais elle pourrait encoder tout autre attribut tel que l'âge ou la catégorie socioprofessionnelle. La loi conditionnelle de $\boldsymbol{R}$, Equation 1, devient :

$$(R_{ij}|U_{iq}V_{jl} = 1, A_i, B_j, G_j, Cov_i) \overset{\text{iid}}{\sim} \mathcal{N}\Big(\pi_{ql} + A_i + B_j + Cov_i G_j, \sigma^2\Big)$$
$$\text{avec } G_j \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_G^2\Big), \ \sigma_G^2 \in \mathbb{R}_+^* \text{ , la variable latente liée à l'objet } j \text{ et } Cov_i \in \{-1, 1\} \ ,$$

la covariable encodant le sexe de l'individu i.

2.2 Modèle de manquement

Le modèle de manquement génère la matrice $\boldsymbol{M}$ « masquant » les données complètes. Nous supposons que chaque utilisateur a une tendance à noter et que de même, chaque objet a sa propre tendance à être noté. Notre modèle de manquement utilise deux variables latentes pour représenter ces effets, que nous appelons ici « propensions ».

Un modèle linéaire mixte modélise la probabilité d'observation à partir de la propension totale, notée P_{ij}, associée à l'utilisateur i et l'objet j. Cette propension est la somme d'un effet fixe, global, noté μ, d'un écart aléatoire associé à l'utilisateur i, noté C_i et d'un écart aléatoire associé à l'objet j, noté D_j. Ainsi, la propension totale est

$$\begin{aligned} & P_{ij} = \mu + C_i + D_j \ , \\ \text{avec } & C_i \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_C^2\Big), \qquad \sigma_C^2 \in \mathbb{R}_+^* \\ & D_j \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_D^2\Big), \qquad \sigma_D^2 \in \mathbb{R}_+^* \ . \end{aligned}$$

La probabilité d'observer la note R_{ij} est obtenue en appliquant une fonction logistique à la propension totale de l'utilisateur i et de l'objet j :

$$p(M_{ij} = 1|C_i, D_j) = \frac{1}{1 + \exp(-P_{ij})} \ . \tag{2}$$

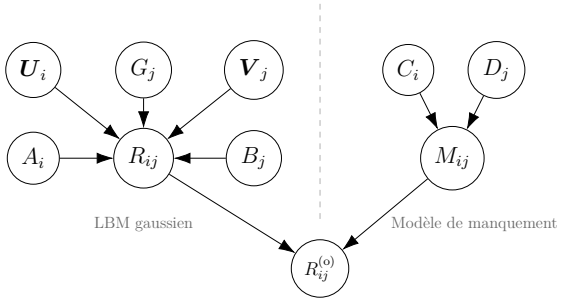


FIGURE 1 – Graphe de dépendances du modèle à blocs latents ordinal couplé au modèle de manquement. $\boldsymbol{R}^{(o)}$ est la matrice observée contenant les notes discrètes. $\boldsymbol{R}$ est la matrice complète de notes continues. $\boldsymbol{M}$ est la matrice « masquant » partiellement les données complètes.

Films prédit comme étant les plus appréciés conditionnellement au fait d'être une femme
Dirty Dancing
Rocky Horror Picture Show
Sound of Music
Grease
Jumpin' Jack Flash

Films prédit comme étant les plus appréciés conditionnellement au fait d'être un homme
The Good, The Bad and The Ugly
Animal House
Caddyshack
Dumb & Dumber
The Exorcist

FIGURE 2 – Films du jeu de données ML-1M dont la valeur inférée de G_j est la plus haute (liste du haut) et la plus basse (liste du bas).

2.3 Modèle joint

Le modèle joint est obtenu en associant le modèle des données complètes et le modèle de manquement. Connaissant la matrice d'évaluations $\boldsymbol{R}$ et la matrice de masque $\boldsymbol{M}$, les éléments de la matrice d'évaluations observées $\boldsymbol{R}^{(o)}$ sont modélisés comme suit :

$$\left(R_{ij}^{(o)} \middle| R_{ij}, M_{ij}\right) = \begin{cases} \sum_{k=1}^{K} k \, \mathbb{1}_{]\zeta_{k-1};\zeta_k]}(R_{ij}) & \text{si} \quad M_{ij} = 1 \\ \text{NA} & \text{si} \quad M_{ij} = 0 \end{cases}$$

$$\text{avec} \quad -\infty = \zeta_0 < \zeta_1 < ... < \zeta_{K-1} < \zeta_K = +\infty$$

des seuils fixés. Un résumé sous forme de graphe de dépendances du modèle joint est présenté Figure 1.

3 Inférence et prédiction

3.1 Critère d'optimisation

La log-vraisemblance du LBM comporte une somme dont le nombre de termes croit exponentiellement avec le nombre de valeurs que peuvent prendre les variables latentes. L'estimateur du maximum de vraisemblance du LBM n'est donc pas calculable, et nous utilisons une inférence variationnelle.

L'inférence variationnelle consiste à introduire q_γ, une distribution sur les variables latentes, dont la forme est choisie de manière à permettre le calcul. L'inférence variationnelle consiste à optimiser une borne inférieure de la log-vraisemblance, à savoir :

$$\begin{aligned}\mathcal{J}(\gamma,\theta) &= \log p\Big(\boldsymbol{R}^{(\mathrm{o})};\theta\Big) - \mathrm{KL}\left(q_\gamma \,\Big\|\, p\Big(L|\boldsymbol{R}^{(\mathrm{o})}\Big)\right) \\ &= \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{R}^{(\mathrm{o})}, L;\theta\Big)\Big] \ . \end{aligned} \tag{3}$$

où KL est la divergence de Kullback–Leibler, $\mathcal{H}$ désigne l'entropie différentielle, $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2, \sigma_G^2)$ est la concaténation de tous les paramètres du modèle, et $L = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{G})$ est la concaténation de toutes les variables latentes.

Le calcul effectif du critère $\mathcal{J}(\gamma,\theta)$ nécessite de restreindre l'espace des distributions variationnelles ; nous choisissons les formes suivantes :

$$\begin{aligned}
&\forall i, \quad \boldsymbol{U}_i|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{M}\Big(1;\boldsymbol{\tau}_i^{(U)}\Big) && \forall j, \quad \boldsymbol{V}_j|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{M}\Big(1;\boldsymbol{\tau}_j^{(U)}\Big) \\
&\forall i, \quad A_i|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{N}\Big(\nu_i^{(A)}, \rho_i^{(A)}\Big) && \forall j, \quad B_j|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{N}\Big(\nu_j^{(B)}, \rho_j^{(B)}\Big) \\
&\forall i, \quad C_i|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{N}\Big(\nu_i^{(C)}, \rho_i^{(C)}\Big) && \forall j, \quad D_j|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{N}\Big(\nu_j^{(D)}, \rho_j^{(D)}\Big) \\
& && \forall j, \quad G_j|\boldsymbol{R}^{(\mathrm{o})} \underset{q_\gamma}{\sim} \mathcal{N}\Big(\nu_j^{(G)}, \rho_j^{(G)}\Big) \ .
\end{aligned}$$

L'indépendance des variables latentes conditionnellement à $\boldsymbol{R}^{(\mathrm{o})}$ simplifie le calcul de $\mathcal{J}(\gamma,\theta)$, et en permet une ré-écriture :

$$\mathcal{J}(\gamma,\theta) = \mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{R}^{(\mathrm{o})}|L\Big)\Big] - \mathrm{KL}\left(q_\gamma \,\|p(L;\theta)\right) \ . \tag{4}$$

L'approximation variationnelle ne suffit toutefois pas à faire le calcul analytique du premier terme de l'équation (4), qui fait intervenir une fonction non linéaire d'une loi normale. Nous approchons numériquement cette espérance en échantillonnant $p(\boldsymbol{R}^{(\mathrm{o})}|L)$ sous la distribution variationnelle. Cette approximation du critère $\mathcal{J}(\gamma,\theta)$ est alors optimisée en une seule étape par l'algorithme Adam [2], qui accélère un gradient stochastique par une estimation des moments de ces gradients.

3.2 Prédicteur des données non observées

Un système de recommandation vise à proposer à chaque utilisateur un ensemble d'objets pouvant l'intéresser. Pour ce faire, les objets doivent être classés par ordre de pertinence en prédisant les notes manquantes $\boldsymbol{R}^{(\mathrm{m})}$. Nous évaluons avec le score de pertinence *Normalized Discounted Cumulative Gain* (NDCG), la recommandation associée à chaque utilisateur. Nous notons $\hat{\boldsymbol{R}}$ la matrice des notes prédites, dont chaque entrée est une note générée par le modèle. En utilisant les maxima *a posteriori* sous la loi variationnelle, nous construisons le prédicteur des données manquantes suivant :

$$\hat{R}_{ij} = \sum_{ql} \tau_{iq}^{(U)} \hat{\pi}_{ql} \tau_{jl}^{(V)} + \nu_i^{(A)} + \nu_j^{(B)} \tag{5}$$

TABLE 1 – Valeur moyenne du score de pertinence Normalized Discounted Cumulative Gain

Modèle	NDCG@5	NDCG@10	NDCG@15
Avec covariable	0.8887	0.9060	0.9289
Sans covariable	0.8905	0.9072	0.9297

Ce prédicteur n'introduit pas l'effet $Cov_i G_j$ lié à la covariable et ainsi, sous condition que les groupes formés par le coclustering soient indépendant du sexe des individus alors l'ordonnancement des objets est lui aussi indépendant du sexe.

4 Expérimentations

Le jeu de données MovieLens [1] 1M regroupe des notes de films qui s'échelonnent de 1 à 5 (5 pour les films les plus appréciés) ainsi que des informations sur le sexe des individus. Dans ce jeu de données réel, les données manquantes $\boldsymbol{R}^{(\mathrm{m})}$ ne sont pas connues ; nous estimons alors la NDCG moyenne sur des ensembles de tests construits par validation croisée (5 blocs) en conservant 20 évaluations par utilisateur. Nous entraînons notre modèle et sa variante, sans covariable, sur ce jeu de donnée avec un nombre de classes en ligne et en colonne fixé à 15.

Afin de mesurer l'indépendance entre le sexe de l'utilisateur et leur appartenance aux groupes, nous calculons la statistique du χ^2 sur le tableau de contingence construit avec les effectifs d'hommes et de femmes dans chaque groupe. La valeur de la statistique du χ^2 pour le modèle sans covariable est de 44.4 (avec une *p-value* de $2.2 \cdot 10^{-4}$) et pour celle du modèle avec covariable est de 18.0 (avec une *p-value* de 0.26). L'utilisation de la covariable permet ainsi de limiter fortement l'effet du sexe sur la structure du coclustering.

Conformément à nos attentes, les valeurs moyennes des NDCG rassemblées dans le tableau 1, présentent de moins bonnes performances pour le modèle avec covariables. Cela est lié au fait que le prédicteur (Equation 5) ignore le sexe de l'utilisateur. Cette baisse de performance est le coût à payer pour obtenir une recommandation équitable entre sexes.

Références

[1] G. GOVAERT et M. NADIF : Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, February 2008.

[2] D. P. KINGMA et J. BA : Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

1. `https://grouplens.org/datasets/movielens/`

MIAMI: Mixed data Augmentation Mixture.

Robin Fuchs [1] & Denys Pommeret [2] & Samuel Stocksieker [3]

[1] *Institut de Mathématiques de Marseille Campus de Luminy Case 901, 163 Avenue de Luminy, Marseille, 13009 Marseille - robin.fuchs@univ-amu.fr*
[2] *ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, France - denys.pommeret@univ-amu.fr*
[3] *Institut de Mathématiques de Marseille Campus de Luminy Case 901, 163 Avenue de Luminy, Marseille, 13009 Marseille - samuel.stocksieker@univ-amu.fr*

Résumé. Les jeux de données mixtes sont composés de variables continues, catégorielles, ordinales, binaires et de comptage. Ce type de données est présent dans toutes les disciplines : sciences sociales, écologie, médecine, astronomie, physique, etc. Lorsque ces données sont coûteuses à acquérir ou concerne des événements rares, peu de procédures existent, en particulier en régression, pour rééchantillonner de tels jeux de données. Ceci est particulièrement dû à la complexité de l'espace dans lequel évolue les observations mixtes. Le présent travail propose d'utiliser une représentation continue de cet espace grâce à un modèle récemment introduit dans la littérature, le M1DGMM, et d'utiliser ce modèle comme processus générateur de nouvelles observations synthétiques. Cette approche peut s'appliquer en classification pour renforcer des classes minoritaires, en régression pour créer des modalités sous-représentées de variable dépendante et/ou de covariables, ou encore en clustering pour générer des combinaisons de variables sous-représentées. Enfin, cette nouvelle approche peut être utilisée pour l'imputation de données manquantes.

Mots-clés. Données déséquilibrées, données mixtes, données synthétiques, données manquantes

Abstract. Mixed data sets are composed of continuous, categorical, ordinal, binary, and count variables. This type of data is present in all disciplines: social sciences, ecology, medicine, astronomy, physics, etc. When these kind of data are expensive to acquire or describe rare events, few procedures exist to resample such datasets, especially in the regression case. This is notably due to the complexity of the space in which the observations evolve. The present work proposes to use a continuous representation of this space learnt by a model recently introduced in the literature, the M1DGMM, and to use this model as a data-generating process. This approach can be applied in classification to reinforce minority classes, in regression to create under-represented modalities of dependent variable and covariates, or even in clustering to generate combinations of under-represented variables. Finally, this new methodology can be applied to impute missing data.

Keywords. Imbalanced data, mixed data, synthetic data, missing data

1 Problem presentation

The proposed method, referred to as MIxed data Augmentation MIxture (MIAMI), is designed to tackle two types of statistical issues: imbalanced datasets and missing data.

A broad definition of imbalanced datasets is used in the following and refers to datasets where some modalities of the dependent variable but also of the covariates are under-represented (these under-represented modalities are actually often the ones of main interest). Imbalanced datasets are common in many Machine Learning tasks such as regression or classification which received greater attention (krawczyk, 2016) or in clustering tasks, a framework less treated in the literature. When data are continuous, well established methods have been developed, often based on interpolation between observations such as in the SMOTE algorithm (Chawla et al., 2002) and its adaptations (e.g. Torgo et al. (2013)), or by applying a perturbation to a given observation as in Lee (2000).

When data are mixed, these methods cannot be used as such, given the more complex topology of mixed datasets spaces. Some adaptations have been proposed such as transforming non-continuous data (e.g. one-hot encoding of categorical variables, considering ordinal variables as continuous etc.) as in Engelmann and Lessmann (2020) or to perform random draws over the modalities of non-continuous variables based on their empirical frequencies (Branco, 2017).

Similarly for missing data imputation, more methods have been developed in the continuous data case compared to the mixed data case. Continuous missing data are often dealt with using standard imputations methods which for instance assign the conditional mean of the variable. Concerning mixed data imputation, several approaches propose a treatment of categorical and ordinal variables in order to apply methods designed for the continuous framework. Audigier et al. (2016) introduced new imputation approaches based on principal component methods offering the possibility to deal with several data types and data sizes. He (2012) also performed dimension reduction using a projection in a Gaussian latent space. Finally, Murray (2017) and Hu et al. (2018) have dealt with missing mixed data using a Bayesian Dirichlet process mixture as latent space.

This work is based on the same idea and perform a projection into a continuous space using the recently introduced M1DGMM, and then use this latent representation to conduct missing data imputation and data augmentation. Our approach will be compared to the previously evoked models to benchmark its performance.

2 Model presentation

The M1DGMM is a multi-layer clustering model introduced by Fuchs et al. (2020). Mixed data are plunged into a continuous space using an extended version of a Generalized Linear Latent Variable Model (GLLVM) (Moustaki and Knott, 2000; Moustaki, 2003) that acts has an embedding layer. The embedded data are then clustered while going through

DGMM layers (Viroli and McLachlan, 2019).
Let Y denote the observed data, n the number of observations, p the number of covariates and the m the number of dependent variables. In a supervised framework, $Y = (X, y)$ with X the covariates and y the dependent variable (potentially multivariate, if $m > 1$). In this case, Y is of dimension $n \times (p + m)$. In an unsupervised framework, $Y = X$ has dimension $n \times p$. Denoting by i the observation index and j the variable index, the model can be written in the following way:

$$\begin{cases} Y_i \to z_i^{(1)} \text{ through GLLVM link via } (\lambda^{(0)}, \Lambda^{(0)}) \\ z_i^{(1)} = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(2)} + u_{i,k_1}^{(1)} \text{ with probability } \pi_{i,k_1}^{(1)} \\ \dots \\ z_i^{(L-1)} = \eta_{k_{L-1}}^{(L-1)} + \Lambda_{k_{L-1}}^{(L-1)} z_i^{(L)} + u_{k_{L-1}}^{(L-1)} \text{ with probability } \pi_{i,k_L}^{(L-1)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases} \tag{1}$$

where the "GLLVM link" stands for the fact that $\forall j \in [1, p]$, $f(Y_j | z^{(1)}, \Theta)$ belongs to an exponential family, with Y_j the jth variable of Y. For example if Y_j is a count variable with support $[1, n_j]$, one can choose the Binomial distribution as link function and obtain:

$$f(Y_j | z^{(1)}, \Theta) = \binom{n_j}{Y_j} f(z^{(1)})^{Y_j} (1 - f(z^{(1)}))^{n_j - Y_j} \tag{2}$$

This example can be adapted for binary variables, by taking $n_j = 1$ and consider a Bernoulli distribution. In the simulations, the ordered multinomial distribution is used for the ordinal data, the unordered multinomial distribution for categorical variable and the Gaussian distribution for continuous variables. In cases where the support of a continuous variable is not $[-\infty, +\infty]$, other distributions than the Gaussian distribution can of course be considered.

The graphical model associated with (1) is given in Figure 1.

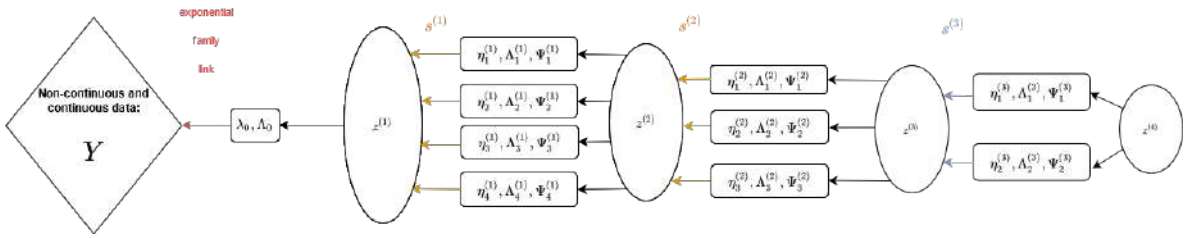


Figure 1: Graphical model of a M1DGMM

3 Data augmentation procedure

3.1 General mechanism

The M1DGMM is first trained in order to learn a continuous representation of the already collected data Y. The variables estimated during the run of the model are denoted by a tilde in the sequel.

Using Bayes rule, one obtains:

$$f(Y|\tilde{\Theta}) = \frac{f(\tilde{z}^{(1)}|\tilde{\Theta})f(Y|\tilde{z}^{(1)},\tilde{\Theta})}{f(\tilde{z}^{(1)}|Y,\tilde{\Theta})} \tag{3}$$

$$\propto f(\tilde{z}^{(1)}|\tilde{\Theta})\prod_{j=1}^{p} f(Y_j|\tilde{z}^{(1)},\tilde{\Theta}). \tag{4}$$

Where (4) comes from the conditional independence assumption which stipulates that the original variables of the dataset are mutually independent with respect to the latent variable, *i.e.* $\forall j, j' \in [1,p]^2$, $Y_j \perp\!\!\!\perp Y_{j'}|z^{(1)}$. Intuitively this assumption, which derived from the GLLVM, means that the latent representation captures all mutual information between the variables and that variable specific information is carried by the GLLVM link function coefficients.

Using (4), the simulations of pseudo-observations Y^* might be performed in the following way:

- Use the M copies of $\tilde{z}^{(1)}|\tilde{\Theta}$ simulated during the M1DGMM training
- Use $\tilde{z}^{(1)}$ to draw some observations from $f(Y_j^*|\tilde{z}^{(1)},\tilde{\Theta})$

The choice of the number of pseudo-observations to simulate can be ruled either by changing M the number of copies of the latent variables $\tilde{z}^{(1)}$ or the number of pseudo-observations Y^* to draw for each $\tilde{z}^{(1)}$.

Using the conditional independence assumption is convenient and seems not to entail performances loss (Fuchs et al., 2020). Yet, the conditional independence between the original variables makes it more difficult to force the model to generate pseudo-observations that present some specific combinations of modalities between variables.
As a result, in order to select only pseudo-observations presenting under-represented modalities, a simple accept-reject procedure is used. The user can specify the ranges in which the pseudo-observations have to belong, the model generates a large number of points and keep those belonging to the pre-specified ranges. The computational cost of this procedure remains affordable as generating more pseudo-observations does not entail

a significant additional computational burden. However, more advanced procedures could be explored in future research, such as using only the draws belonging to interesting latent space areas to simulate Y^* from.
The MIAMI method refers to the generation of pseudo-observations by M1DGMM coupled with the accept-reject procedure.

3.2 Applications

The pseudo-observations Y^* generated by MIAMI can be used in different applications:

Imbalanced data

- In imbalanced regression problems, synthetic dependent variables or covariates can be obtained by generating a matrix $Y^* = (X^*, y^*)$, where X^* stands for the design matrix of the regression model, and Y^* is the dependent variable (potentially multivariate). Then, both X^* and Y^* are simultaneously selected. The accept-reject procedure can be used symmetrically: either depending of the value of X^* in the case of imbalanced design, or depending of the value of y^* in the case of imbalanced labels.
- In classification problems where $Y^* = (X^*, y^*)$ or in clustering problems where $Y^* = X^*$, the criteria of selection depends on the representation of the minority classes. The algorithm is run as long as these classes are not sufficiently re-balanced.
- In "comparison" problems, such as clinical trials with comparison of cohorts, the algorithm can be used to achieve a fixed ratio of observations, as for instance treatment versus placebo. The algorithm can create synthetic patients such that the number of patients in a group is proportional to the number of patients in another group.

Missing data The algorithm is initialized using simple imputations methods (e.g. imputation by the conditional mean of the group). Then, MIAMI is run several times, iteratively replacing the missing values by new estimates determined by the pseudo-observations matching the characteristics of incomplete observations.

3.3 Numerical illustration

The method will be illustrated by performing simulations on several practical cases.

Bibliographie

References

Audigier, V., F. Husson, and J. Josse (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification 10*, 5–26.

Branco, Torgo, R. (2017). Smogn: a pre-processing approach for imbalanced regression. *Proceedings of Machine Learning Research 74*, 36—50.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Engelmann, J. and S. Lessmann (2020). Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *arXiv preprint arXiv:2008.09202*.

Fuchs, R., D. Pommeret, and C. Viroli (2020). Mixed data deep gaussian mixture model: A clustering model for mixed datasets. *arXiv preprint arXiv:2010.06661*.

He (2012). *Multiple Imputation of High-dimensional Mixed Incomplete Data*. Ph. D. thesis, University of California.

Hu, J., J. P. Reiter, Q. Wang, et al. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis 13*(1), 183–200.

krawczyk (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*(4), 221—232.

Lee, S. S. (2000). Noisy replication in skewed binary classification. *Computational statistics & data analysis 34*(2), 165–191.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology 56*(2), 337–357.

Moustaki, I. and M. Knott (2000). Generalized latent trait models. *Psychometrika 65*(3), 391–411.

Murray, R. (2017). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association 111*(516), 1466—1479.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). Smote for regression. In *Portuguese conference on artificial intelligence*, pp. 378–389. Springer.

Viroli, C. and G. J. McLachlan (2019). Deep gaussian mixture models. *Statistics and Computing 29*(1), 43–51.

Analyse bayésienne des modèles de médiation et de modération

Jean-Michel GALHARRET [1] & Anne PHILIPPE [1]

[1] *Laboratoire Jean Leray , 2 rue de la Houssinière, 44322 Nantes Cedex 3 , jean-michel.galharret@univ-nantes.fr*

Résumé. En sciences humaines, l'analyse de médiation consiste à étudier si l'effet d'une variable d'exposition X sur une variable réponse Y peut être décomposé en un effet direct et un effet indirect via une troisième variable M (le médiateur). Une question additionnelle est de savoir si ces effets sont les mêmes dans deux populations données. Il s'agit alors d'étudier un modèle de médiation modérée. Nous proposons une extension des g-priors utilisés en régression linéaire aux modèles de médiation et de médiation modérée. Ce choix de lois a priori permet d'obtenir une forme explicite de la distribution marginale et ainsi du facteur de Bayes. Celui-ci servira à tester l'existence des effets indirects en médiation et l'existence d'une modération. Nous appliquons cette méthodologie à une étude en psychologie du travail sur le lien entre le leadership habilitant et l'engagement organisationnel.

Mots-clés. Facteur de Bayes, $g-$priors, médiation, modération

Abstract. In social sciences, mediation analysis consists of studying whether the effect of an exposure variable X on a response variable Y can be decomposed into a direct effect and an indirect effect via a third variable M. An additional question is whether these mediated effects are the same in two given populations. In this case, a model of moderate mediation should be studied. We propose an extension of the g-priors to the mediation and moderate mediation models. This prior distribution provides explicit forms for the marginal distribution and thus for the Bayes factor. This will be used to test the existence of indirect effects in mediation and the existence of moderation. We apply this methodology to a study in psychology of the relation between ewpowering leadership and organizational commitment.

Keywords. Bayes factor, $g-$priors, mediation, moderation

1 Introduction

Les modélisations proposées ont été motivées par une étude visant à comprendre le lien entre les composantes du leadership habilitant (i.e. le développement de la confiance au travail, de l'autonomie, de la participation aux décisions et du sens du travail) et l'engagement organisationnel (i.e. envers l'entreprise dans laquelle travaille l'employé) à

travers leurs liens au bien-être au travail (médiateur). Un enjeu de cette étude est de voir si les liens observés sont les mêmes dans deux structures différentes (pompiers et employés de l'industrie).

Pour la première partie de l'étude on estime les paramètres $a, c \in \mathbb{R}^p,\ i_2, i_1, b, d_0 \in \mathbb{R}$ dans le système d'équations linéaires suivant :

$$[\mathcal{M}_0] : \begin{cases} Y = i_2 + c^T X + bM + e_0 W + \varepsilon_Y \\ M = i_1 + a^T X + d_0 W + \varepsilon_M \end{cases} \tag{1}$$

où $Y \in \mathbb{R}^n$ est une variable réponse continue (ici l'engagement organisationnel), $M \in \mathbb{R}^n$ est un médiateur continu (le bien-être au travail), $X \in \mathbb{R}^{n,p}$ est la matrice des variables d'exposition (ici les quatre composantes du leadership habilitant) et W est ici binaire et correspond aux deux organisations. On suppose que X et W sont déterministes et que les termes d'erreurs sont gaussiens $\varepsilon_M \sim \mathcal{N}_n(0, \sigma_M^2 I_n)$, $\varepsilon_Y \sim \mathcal{N}_n(0, \sigma_Y^2 I_n)$.

On adopte les définitions des effets directs et indirects de chaque composante X_k introduits par Hayes, 2018. L'effet direct de X_k sur Y est égal c_k et son effet indirect via M est le produit $a_k b$. Tester l'existence d'un effet direct est un problème classique. Pour l'effet indirect, l'approche standard consiste à approcher la loi du produit $a_k b$ par bootstrap ou par une loi gaussienne grâce à la $\Delta-$méthode (voir MacKinnon et al. 2002).

Tester la modération des effets obtenus dans $[\mathcal{M}_0]$ consiste à ajouter des termes d'interaction dans les deux équations de régression. Ces termes seront notés $X : W$ pour l'interaction entre les X_k et W. On choisit de modérer les liens X, M et M, Y ce qui donne (voir aussi Figure 1) :

$$[\mathcal{M}_1] : \begin{cases} Y = i_2 + c^T X + bM + e_0 W + e_1^T X : W + f_1 M : W + \varepsilon_Y, \\ M = i_1 + a^T X + d_0 W + d_1^T X : W + \varepsilon_M \end{cases} \tag{2}$$

où $e_1, d_1 \in \mathbb{R}^p,\ f_1 \in \mathbb{R}$ sont les effets modérés (appelés aussi effets conditionnels voir Preacher et al. 2007 pour plus de détails). Tester l'existence de la modération des effets revient donc à tester :

$$\mathcal{H}_0 : d_1 = e_1 = 0_p,\ f_1 = 0. \tag{3}$$

Les modèles emboités $[\mathcal{M}_0]$ et $[\mathcal{M}_1]$ sont alors comparés en utilisant le test de rapport de varisemblance (LR-test)

$$\lambda = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta}_1)}{\mathcal{L}(\widehat{\theta}_0)} \right),$$

où $\widehat{\theta}_1$ et $\widehat{\theta}_0$ sont respectivement les estimateurs du maximum de vraisemblance de $[\mathcal{M}_0]$ et $[\mathcal{M}_1]$. Sous $\mathcal{H}_0$ λ est distribuée selon une loi du χ^2 à $2p+1$ degrés de liberté.

Dans ce travail, nous adaptons les $g-$priors aux modèles de médiation. L'intérêt est d'obtenir des formes explicites du facteur de Bayes pour tester l'existence de ces différents effets.

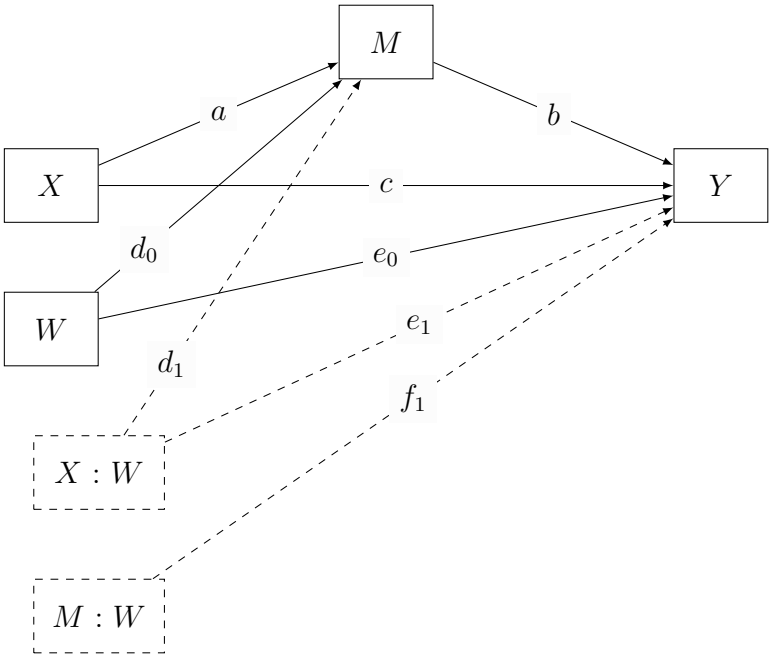


Figure 1: Modèle de médiation (lignes pleines) et modèle de médiation modérée (avec l'ajout des lignes en pointillés).

2 Extension des g-priors de Zellner au modèle de médiation et médiation modérée

Les $g-$priors ont été introduit par Zellner (1984) pour les coefficients des modèles de régression multiple $Y = \mathbb{X}\beta + \varepsilon$, où $\mathbb{X} \in \mathbb{R}^{n\times p}$ est la matrice de design, $Y \in \mathbb{R}^n$ le vecteur réponse, et l'erreur ε qui est supposée gaussienne $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Le $g-$priors sur les paramètres (β, σ^2) est

$$\beta|\sigma^2 \sim \mathcal{N}_p(\tilde{\beta}, g\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}) \text{ and } \pi(\sigma^2) \propto \sigma^{-2}. \tag{4}$$

noté par la suite $(\beta, \sigma^2) \sim \mathcal{Z}_g(g, \widetilde{\beta}, \mathbb{X})$.
Soit $\theta := (\theta_M, \sigma^2_M, \theta_Y, \sigma^2_Y)$ le vecteur des paramètres du modèle de médiation défini par (1), avec $\theta_M = (i_1, a, d_0)$ et $\theta_Y = (i_2, c, b, d_0)$. Dans le modèle global (1), la distribution jointe de M, Y, θ se décompose sous la forme

$$p(m, y, \theta|X, W) = f_Y(y|m, \theta, X, W) f_M(m|\theta, X, W)\pi(\theta),$$

où f_Y est la densité de la loi $\mathcal{N}(i_2 + c^T X + bm + e_0 W, \sigma^2_Y)$ et f_M celle de la loi $\sim \mathcal{N}(i_1 + a^T X + d_0 W, \sigma^2_M)$. Une solution est de supposer que les paramètres θ_M, θ_Y sont indépendants, c'est à dire écrire $\pi(\theta) = \pi_M(\theta_M, \sigma^2_M)\pi_Y(\theta_Y, \sigma^2_Y)$. Cependant, cette solution

ne permet pas d'utiliser les $g-$priors car M apparaît dans la matrice de désigne de π_Y. Nous proposons de décomposer la distribution jointe de M, Y, θ sous la forme

$$\begin{aligned} p(m, y, \theta_Y, \sigma_Y^2, \theta_M, \sigma_M^2 | X, W) &= f_Y(y | M = m, \theta_Y, \sigma_Y^2, \theta_M, \sigma_M^2, X, W) \times \\ &\quad \pi(\theta_Y, \sigma_Y^2 | M = m, \theta_M, \sigma_M^2, X, W) \times \\ &\quad f_M(m | \theta_M, \sigma_M^2, X, W) \pi(\theta_M, \sigma_M^2 | X, W). \end{aligned}$$

qui compte tenu des dépendances décrites dans le DAG Figure 2 d'obtenir

$$\begin{aligned} p(m, y, \theta_Y, \sigma_Y^2, \theta_M, \sigma_M^2 | X, W) &= f_Y(y | M = m, \theta_Y, \sigma_Y^2, X, W) \pi_Y(\theta_Y, \sigma_Y^2 | M = m, X, W) \times \\ &\quad f_M(m | \theta_M, \sigma_M^2, X, W) \pi_M(\theta_M, \sigma_M^2 | X, W), \end{aligned} \tag{5}$$

On choisit pour π_M et π_Y les $g-$priors pour les matrices de design $[\mathbb{1}, X, W]$ et $[\mathbb{1}, X, W, M]$ respectivement. En l'absence d'information sur θ_M, θ_Y, on choisit $\widetilde{\theta}_M = \widetilde{\theta}_Y = 0$ et $g_M = g_Y = n$ qui sont les choix standards (voir par exemple Marin et Robert 2014). Le modèle bayésien est alors complètement défini par (5) et ce choix de $g-$priors et la loi a posteriori est :

$$\pi(\theta_M, \sigma_M^2, \theta_Y, \sigma_Y^2 | X, W, M, Y) \propto p(M, Y, \theta_Y, \sigma_Y^2, \theta_M, \sigma_M^2 | X, W). \tag{6}$$

On déduit facilement de (6) les lois a posteriori des effets directs et indirects. Biesanz et al. 2010 proposent de tester l'existence de ces effets en utilisant leurs intervalles de crédibilité.

3 Facteur de Bayes pour tester le modèle de médiation modérée

Tester si W modère les liens entre X, M, Y peut être vu comme un problème de sélection de modèle entre $[M_0]$ et $[M_1]$. Comme précédemment, nous utilisons les $g-$priors pour chacun des deux modèles, c'est à dire :

- Modèle $\mathcal{M}_1$:

$$\begin{aligned} Y &= i_2 + c^T X + bM + e_0 W + e_1^T X : W + f_1 M : W + \varepsilon_Y, \\ (\theta_Y, \sigma_Y^2) | X, W, M &\sim \mathcal{Z}_g(g_2, \widetilde{\beta}_2, \mathbb{X}_{Y_1}) \text{ with } \mathbb{X}_{Y_1} = \left[\mathbb{1}, W, X, M, X : W, M : W\right] \\ M &= i_1 + a^T X + d_0 W + d_1^T X : W + \varepsilon_M \\ (\theta_M, \sigma_M^2) | X, W &\sim \mathcal{Z}_g(g_1, \widetilde{\beta}_1, \left[\mathbb{1}, W, X, X : W\right]) \text{ with } \mathbb{X}_{M_1} = \left[\mathbb{1}, W, X, X : W\right] \end{aligned}$$

- Modèle $\mathcal{M}_0$:

$$\begin{aligned} Y &= i_2 + c^T X + bM + e_0 W + \varepsilon_Y, \\ (\theta_Y, \sigma_Y^2) | X, W, M &\sim \mathcal{Z}_g(g_2, \check{\beta}_2, \left[\mathbb{1}, W, X, M\right]) \text{ with } \check{\beta}_2 = (\widetilde{\beta}_2^1, ..., \widetilde{\beta}_2^{p+3}) \text{ and } \mathbb{X}_{Y_0} = \left[\mathbb{1}, W, X, M\right] \\ M &= i_1 + a^T X + d_0 W + \varepsilon_M \\ (\theta_M, \sigma_M^2) | X, W &\sim \mathcal{Z}_g(g_1, \check{\beta}_1, \left[\mathbb{1}, W, X\right]) \text{ with } \check{\beta}_1 = (\widetilde{\beta}_1^1, ..., \widetilde{\beta}_1^{p+2}) \text{ and } \mathbb{X}_{M_0} = \left[\mathbb{1}, W, X\right] \end{aligned}$$

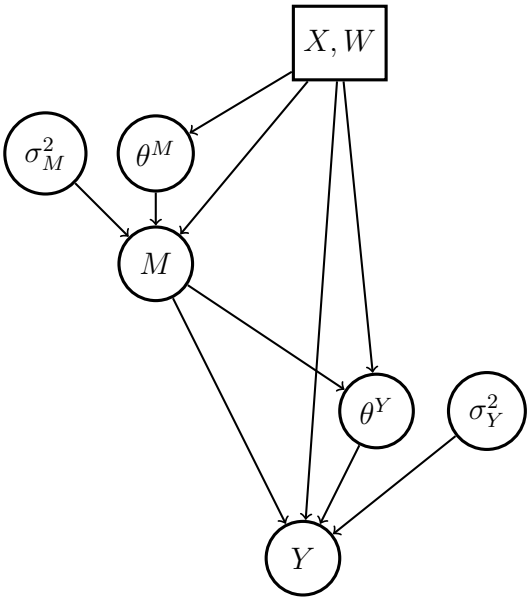


Figure 2: Directed acyclic graph for the mediation model $[\mathcal{M}_0]$.

Compte tenu de la décomposition de la loi marginale des paramètres on obtient la forme explicite pour le facteur de Bayes.

Proposition 1 *Le facteur de Bayes pour comparer les modèles $\mathcal{M}_1$ versus $\mathcal{M}_0$ définis ci-dessus est donnée par*

$$BF_{10} = \frac{(g_1+1)^{p/2}}{(g_2+1)^{(p+1)/2}} \left(\frac{Y^TY - \frac{g_2}{g_2+1} Y^T\mathbb{X}_{Y_0}(\mathbb{X}_{Y_0}^T\mathbb{X}_{Y_0})^{-1}\mathbb{X}_{Y_0}^TY - \frac{\|\mathbb{X}_{Y_0}\check{\beta}_2\|^2}{g_2+1}}{Y^TY - \frac{g_2}{g_2+1} Y^T\mathbb{X}_{Y_1}(\mathbb{X}_{Y_1}^T\mathbb{X}_{Y_1})^{-1}\mathbb{X}_{Y_1}^TY - \frac{\|\mathbb{X}_{Y_1}\widetilde{\beta}_2\|^2}{g_2+1}} \right)^{n/2} \times$$
$$\left(\frac{M^TM - \frac{g_1}{g_1+1} M^T\mathbb{X}_{M_0}(\mathbb{X}_{M_0}^T\mathbb{X}_{M_0})^{-1}\mathbb{X}_{M_0}^TM - \frac{\|\mathbb{X}_{M_0}\check{\beta}_1\|^2}{g_1+1}}{M^TM - \frac{g_1}{g_1+1} M^T\mathbb{X}_{M_1}(\mathbb{X}_{M_1}^T\mathbb{X}_{M_1})^{-1}\mathbb{X}_{M_1}^TM - \frac{\|\mathbb{X}_{M_1}\widetilde{\beta}_1\|^2}{g_1+1}} \right)^{n/2}.$$

On décide que les effets sont modérés lorsque $BF_{10} > 1$. Nous comparons deux approches pour quantifier l'évidence de la décision :

- Nous adaptons à la médiation l'approche proposée par Zhou et Guan 2018 pour la régression linéaire multiple. L'idée "fréquentiste" est d'approcher la distribution de BF_{10} sous l'hypothèse $\mathcal{H}_0$ défini en (3) par un bootstrap paramétrique puis d'en déduire une $p-$value.

- Nous construisons une version alternative basée sur la loi prédictive du modèle bayésien sous $\mathcal{M}_1$. On considère $\widetilde{BF}_{10}$ le facteur de Bayes calculé sur un échantillon

$(\widetilde{Y}, \widetilde{M})$ distribué suivant la loi prédictive sous $\mathcal{M}_1$ On quantifie l'évidence de $\mathcal{M}_1$ contre $\mathcal{M}_0$ par la probabilité de l'événement $\widetilde{BF}_{10} > 1$.

4 Application à l'étude du leadership habilitant

Les différentes méthodes précédentes sont appliquées à l'estimation des effets directs et indirects du leadership habilitant et à la modération par le type d'organisation. Tous les résultats obtenus sont cohérents avec les approches fréquentistes utilisées dans l'article initial de Caillé et al. (2020).

Bibliographie

Biesanz, J. C., Falk, C. F., and Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. Multivariate Behavioral Research, 45:661–701.

Caillé, A., Courtois, N., Galharret, J.-M., and Jeoffrion, C. (2020). Influence du leadership habilitant sur le bien-être au travail et l'engagement organisationnel : étude comparative entre une organisation habilitante et une organisation classique. Psychologie du Travail et des Organisations, 26(3):247 – 261.

Hayes, A. F. (2018). Introduction to Mediation, Moderation, and Conditional Process Analysis. A Regression-based Approach. Methodology in the Social Sciences. Guilford Press.

MacKinnon, D. P., Lockwood, c. M., Hoffman, J. M., West, s. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. Psychological Methods, 7, 83-104.

Marin, J.-M. and Robert, C. (2014). Bayesian essentials with R. Springer Textbooks in Statistics. Springer Verlag, New York.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research, 42, 185-227.

Zellner, A. (1984). Basic Issues in Econometrics. University of Chicago Press.

Zhou, Q. and Guan, Y. (2018). On the null distribution of bayes factor in linear regression. Journal of the American Statistical Association, 113(523):1362–1371.

Propagation of epistemic uncertainties and global sensitivity analysis in seismic risk assessment

Clément Gauchy

In seismic probabilistic risk assessment (SPRA), one wants to quantify the failure of a structural system when subjected to seismic ground motions. Nowadays, regulation authorities prescribes industrials to penalize their modelizations by injecting uncertainties on the model parameters itself. These uncertainties are called epistemic: they can be reduced with a cost corresponding to a certain effort (gathering more data, eliciting expert's knowledge...). This effort could be interpreted as a gain in knowledge of the mechanical model studied. The Uncertainty Quantification (UQ) framework (De Rocquigny et al. [2008]) is well adapted to perform SPRA with epistemic uncertainties. In structural safety, the most studied quantity of interest is the so-called fragility curve of the structure, which is the conditional probability of failure given a specific intensity level of the seismic ground motion. It is possible to propose a global sensitivity analysis of the fragility curve with respect to the uncertainties on the mechanical model parameters, as it is a functional risk curve discussed in Iooss and Le Gratiet [2019]. A Gaussian Process regression on the mechanical response in the same fashion as in Iooss and Le Gratiet [2019] has been proposed to alleviate the computational burden of the sensitivity indices estimation such as aggregated Sobol indices, a natural extension of Sobol indices in the functional case. Sensitivity analysis on scalar quantities of interest derived from the fragility curve is also performed. For instance, given a probability distribution for the intensity level of the seismic ground motion, our quantity of interest will become the seismic probability of failure of the mechanical structure given by the integration of the fragility curve with respect to this probability distribution. Despite its advantages, the Sobol indices measure the influence of the model parameters on the quantity of interest's variance. Recently new types of indices, called moment-independent indices and dedicated to the measure of the influence of the input parameters on the whole probability distribution of the output of interest, have been investigated Veiga [2015]. The maximum mean discrepancy (MMD) based sensitivity indices introduced in da Veiga [2021] are a specific moment-independent indices with appealing properties: it has an ANOVA-like decomposition as the Sobol indices, allowing for simple interpretation, and the same computational cost of estimation as the Sobol indices. However, a kernel function has to be chosen prior to the estimation of these indices.

We propose an original method for choosing an appropriated kernel function based of some theoretical properties of the MMD (Sriperumbudur et al. [2010]): the MMD distance estimator between two empirical distributions $\widehat{\mathbb{P}}_n$ and $\widehat{\mathbb{Q}}_n$ based on a n sized sample of probability measures $\mathbb{P}$ and $\mathbb{Q}$ actually raises the L^2 distance between kernel density estimators of these two distributions. This gives direct interpretation of the MMD based sensitivity

indices.

References

Sébastien da Veiga. Kernel-based anova decomposition and shapley effects – application to global sensitivity analysis, 2021.

Etienne De Rocquigny, Nicolas Devictor, and Stefano Tarantola. *Introducing the Common Methodological Framework*, chapter 1, pages 1–19. John Wiley & Sons, Ltd, 2008. ISBN 9780470770733. doi: https://doi.org/10.1002/9780470770733.ch1. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470770733.ch1`.

Bertrand Iooss and Loïc Le Gratiet. Uncertainty and sensitivity analysis of functional risk curves based on gaussian processes. *Reliability Engineering & System Safety*, 187:58–66, 2019. Sensitivity Analysis of Model Output.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL `http://jmlr.org/papers/v11/sriperumbudur10a.html`.

Sebastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

Active learning strategy for fragility curve estimation using adaptive importance sampling

Clément Gauchy [12] & Cyril Feau [1] & Josselin Garnier [2]

[1] *Universite Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191, Gif-sur-Yvette, France, E-mail: clement.gauchy@cea.fr, cyril.feau@cea.fr*
[2] *CMAP, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France, E-mail: josselin.garnier@polytechnique.edu*

Résumé. Les études probabilistes de sûreté sismiques consiste à évaluer les probabilités de défaillance de structures mécaniques soumises à des excitations sismiques. Ces études nécessitent l'estimation de courbes de fragilité sismique, qui sont la probabilité de défaillance de la structure conditionnellement à une mesure d'intensité du signal sismique. Cependant, leur estimation requiert de nombreuses expériences numériques qui peuvent être très coûteuses en temps calcul, ce qui rend l'estimation par une méthode Monte Carlo inappropriée. L'approche proposé dans ce papier est de réduire la taille de l'échantillon tout en conservant de bonnes perfomances en termes de variance d'estimation grâce à l'échantillonnage préférentiel.

Mots-clés. Echantillonnage préférentiel, apprentissage actif, courbe de fragilité.

Abstract. Seismic probabilistic risk assessment studies consist in evaluating the probabilities of failure of mechanical structures when submitted to seismic ground motions. These studies are often concentrated on fragility curve estimation. The fragility curve is the probability of failure of the structure conditionally to a seismic intensity measure. However, its estimation requires computer experiments involving huge computation time. Such a computational burden makes crude Monte Carlo methods untractable, fragility curves estimation must then be economical in terms of sample size. The proposed approach is to use importance sampling to reduce sample size and to guarantee good performances.

Keywords. Importance sampling, active learning, fragility curves.

1 Introduction

Fragility curves estimation for seismic probabilistic risk assesment (SPRA) or probabilistic based earthquake engineering (PBEE) consists in evaluating structural reliability using a set of seismic ground motions, characterized by their so-called intensity measure (IM) e.g. peak ground acceleration, spectral acceleration,... In our case, the quantity of interest is the fragility curve. It is the conditional probability $\mathbb{P}(Y > C|IM = a)$ that a mechanical

demand Y exceeds a threshold C for a given seismic load a. The mechanical demand Y is obtained through numerical simulations whose computational time is cumbersome. To reduce the computational burden, our methodology consists to interwind importance sampling and statistical learning in an adaptive fashion, in order to reduce the asymptotic variance of the training loss. The paper is structured as follows: Section 2 introduces the statistical learning setting of fragility curve estimation, Section 3 presents Adaptive Importance Sampling (AIS) method applied to fragility curve estimation, Section 4 is dedicated to a real case industrial example concerning the fragility curve of a safety water supply pipe section to a steam generator of a nuclear reactor.

2 Statistical learning framework for fragility curve estimation

In the peculiar setting of seismic probabilistic risk assessment, the input variable is the logarithm of a seismic intensity measure $X = \log(IM)$ so that $\mathcal{X} = \mathbb{R}$ and the output variable is a label $S \in \mathcal{S} = \{0, 1\}$. $(S = 0)$ indicates that the structure resists the seismic excitation, $(S = 1)$ corresponds to a failure state of the structure, basically the seismic excitation induces critical damages that compromise the safety of the structure. Fragility curve is then defined by the conditional expectation $\mu : x \to \mathbb{E}[S|X = x]$. Moreover, the structure of the joint distribution of (X, S) is:

$$\begin{array}{rcl} P(dx, ds) & = & p(ds|x)p(x)dx, \\ p(ds|x) & = & \mu(x)\delta_1(ds) + (1 - \mu(x))\delta_0(ds) \ . \end{array} \tag{1}$$

Remark δ_j is the Dirac distribution at j, $x \to p(x)$ is the probability density of X and $S|X$ is supposed to follow a Bernoulli distribution with parameter $\mu(X)$. Supervised statistical learning aims at building a predictive model of the label S from X using a sample of observed pairs $(X_i, S_i)_{i=1}^n$. In our case, fragility curve estimation aims at learning $x \to \mu(x)$. A classical parametric model for fragility curve estimation is the lognormal model

$$f_\theta(x) = \Phi\left(\frac{x - \log(\alpha)}{\beta}\right) \ , \tag{2}$$

where $\theta = (\alpha, \beta)$. The parameters are estimated using the L_2 loss $\ell(s, f_\theta(x)) = (s - f_\theta(x))^2$.

$$\begin{array}{rcl} R(\theta) & = & \mathbb{E}[(S - f_\theta(X))^2] \\ & = & \mathbb{E}[\mu(X)(1 - \mu(X))] + \mathbb{E}[(\mu(X) - f_\theta(X))^2] \ . \end{array} \tag{3}$$

According to empirical risk estimation, the estimator $\widehat{\theta}_n$ is computed as follows from the sample $(S_i, X_i)_{i=1}^n$ assumed to be i.i.d. with the distribution P:

$$\begin{array}{rcl} \widehat{\theta}_n & = & \mathrm{argmin}_{\theta \in (0,+\infty)^2} \widehat{R}_n(\theta), \\ \widehat{R}_n(\theta) & = & \frac{1}{n}\sum_{i=1}^{n}(S_i - f_\theta(X_i))^2 \ . \end{array} \tag{4}$$

3 Adaptive importance sampling

The use of importance sampling as an active learning strategy has been suggested by Chu et al. [2011]. The main idea is to replace the marginal distribution of X (which is assumed to have a probability density p) into an instrumental density q, and to multiply the loss function by the likelihood ratio to recover an unbiased estimate of the fragility curve. When the instrumental density q favors region of $\mathcal{X}$ where the loss function is large, the empirical risk variance can be drastically reduced. The IS empirical risk is:

$$\check{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)}(S_i - f_\theta(X_i))^2 \ , \tag{5}$$

where $(S_i, X_i)_{i=1}^n$ is assumed to be i.i.d. with the distribution $\check{P}(dx, ds) = p(ds|x)q(x)dx$. Remark that the instrumental density only concern the variable X. The distribution of S conditional to X remains unchanged. We can use this setting for fragility curve estimation, indeed the dataset generation $(X_i, S_i)_{\{1,\dots,n\}}$ is entirely numerical, a modulated and filtered white-noise process is used to simulate synthetic seismic ground motion to retrieve X_i, and a mechanical numerical simulation to retrieve the failure state S_i of the structure studied given the artificial seismic signal. In this setting, synthetic seismic generation is considered cheap in terms of computational resources compared to the mechanical simulation. We thus have a perfect knowledge of p the marginal probability density of X. We will choose the instrumental density q which minimizes the variance of the IS empirical risk $\check{R}_n(\theta)$:

$$q_\theta^* = \operatorname*{argmin}_q \operatorname{Var}(\check{R}_n(\theta)) = \operatorname*{argmin}_q \frac{V_\theta}{n} \ ,$$
$$V_\theta = \int_{\mathcal{X}}\int_{\mathcal{S}} \frac{p(x)^2}{q(x)}(s - f_\theta(x))^4 p(ds|x)dx - R(\theta)^2 \ .$$

Denoting $\tilde{\ell}(x, f_\theta) = \mathbb{E}[(S - f_\theta(X))^4 | X = x] = \mu(x)(1 - f_\theta(x))^4 + (1 - \mu(x))f_\theta(x)^4$, V_θ can be written as:

$$V_\theta = \int_{\mathcal{X}} \frac{p(x)^2}{q(x)}\tilde{\ell}(x, f_\theta(x))dx - R(\theta)^2 \ .$$

The optimal sampling density is of the form:

$$q_{\theta,\mu}^*(x) \propto \sqrt{\tilde{\ell}(x, f_\theta)}p(x) \ .$$

We remark that $q_{\theta,\mu}^*$ depends on the parameter θ of the parametric fragility curve and the true fragility curve μ. We then define an approximation of the optimal sampling density as follows

$$q_\theta(x) \propto \sqrt{f_\theta(x)(1 - f_\theta(x))^4 + (1 - f_\theta(x))f_\theta(x)^4}p(x) \, ,$$

where the unknown true fragility μ is replaced by f_θ. A natural heuristic is to build an adaptive, sequential estimation scheme in which the previous estimate of θ is used to

sample new data pair(s), which in turn can be used to update the current estimation. However, the instrumental density can dramatically increase the variance if the tails of p are overestimated (Owen and Zhou [2000]), and this may happen in particular in the early steps of the estimation scheme. A defensive strategy first used by Hesterberg [1995] tackles this issue, it consists in introducing a mixing parameter $\varepsilon \in (0,1)$ such that the instrumental density is:

$$q_{\theta,\varepsilon}(x) = \varepsilon p(x) + (1-\varepsilon) q_\theta(x) \ .$$

The initial estimator $\tilde{\theta}_0$ is chosen with an heuristic presented in the next section. The full estimation procedure is presented in Algorithm 1.

Algorithm 1 Adaptive Importance Sampling (AIS)

1. Initialization: $\tilde{\theta}_0$
2. For $i = 1, \ldots, n$:
   (a) Sample X_i from $q_{\tilde{\theta}_{i-1},\varepsilon}$
   (b) Call the mechanical simulation at point X_i to get label S_i
   (c) Compute

$$\tilde{\theta}_n = \underset{\theta}{\operatorname{argmin}}\, \tilde{R}_n(\theta), \qquad \tilde{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)}{q_{\tilde{\theta}_{i-1},\varepsilon}(X_i)} \ell(S_i, f_\theta(X_i)). \tag{6}$$

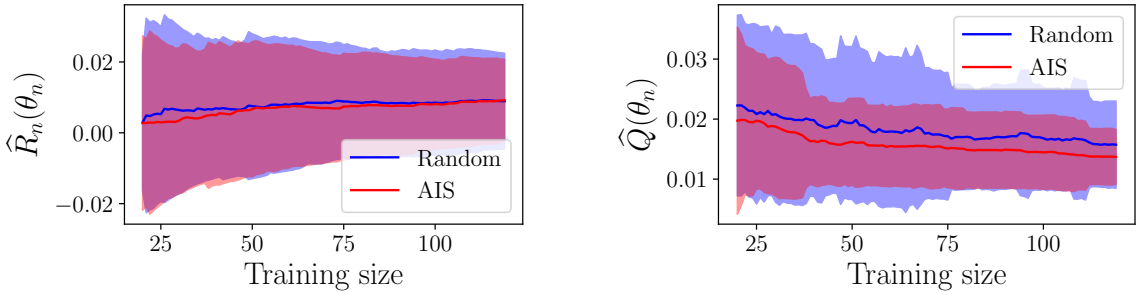
The convergence properties of the estimator $\tilde{\theta}_n$ is assessed in Theorem 1. The proof is based on results of Vaart [1998] and Hall et al. [2014].

Theorem 1 (Consistency and asymptotic normality of $\tilde{\theta}_n$) *Let $X_i, S_i, \tilde{\theta}_i$ be defined as in Algorithm 1. Denote $\psi_\theta(x,s) = \nabla_\theta \ell(s, f_\theta(x))$. Assume that the following conditions are satisfied:*

1. *$\theta_* = \operatorname{argmin}_\theta R(\theta)$ exists and is unique, with $R(\theta)$ defined by (3),*
2. *$\exists \eta > 0$ such that $\sup_{\theta\in\Theta} \iint ||\frac{p(x)\psi_{\theta_*}(x,s)}{q_\theta(x)}||^{2+\eta} P(dx,ds) < +\infty$,*
3. *$\sup_{\theta\in\Theta} \iint \frac{p(x)||\dot{\psi}_{\theta_*}(x,s)\dot{\psi}_{\theta_*}(x,s)^T||}{q_\theta(x)} P(dx,ds) < +\infty$,*
4. *for $\theta \in \mathcal{B}$, with $\mathcal{B}$ a neighborhood of θ_*, $\forall x, s$ $\sup_{\theta\in\Theta} \frac{p(x)\ddot{\psi}_\theta(x,s)}{q_\theta(x)} < +\infty$.*

Then $\tilde{\theta}_n$ is weakly consistent and $\sqrt{n}(\tilde{\theta}_n - \theta_*)$ is asymptotically normal with mean zero and covariance matrix $\ddot{R}(\theta_*)^{-1} V(q_{\theta_*}) (\ddot{R}(\theta_*)^{-1})^T$ with

$$V(q_{\theta_*}) = \mathbb{E}\left[\frac{p(X)}{q_{\theta_*}(X)} (S - f_{\theta_*}(X))^2 \nabla f_{\theta_*}(X) \nabla f_{\theta_*}(X)^T\right].$$



(a) Training error

(b) Test error

Figure 1: Numerical benchmark of the ASG piping system. AIS is initialized with 20 datapoints sampled from $q_{\tilde{\theta}_0,\varepsilon}$. Training error is equal to $\tilde{R}_n(\tilde{\theta}_n)$ for AIS (in red) and $\widehat{R}_n(\widehat{\theta}_n)$ for random sampling (in blue). Test error is equal to $\widehat{Q}(\tilde{\theta}_n)$ for AIS (in red) and $\widehat{Q}(\widehat{\theta}_n)$ for random sampling (in blue)

4 Numerical application: fragility curve estimation of a safety water pipe supply section

The following test case corresponds to the ASG piping system which is a simplified part of a secondary line of a French Pressurized Water Reactor. The main results are outlined in the reference Touboul et al. [1999]. In the following, the random variable R corresponds to the maximum of the out-of-plane rotation of the elbow of the pipe. The Bernoulli variable which indicates the failure state is defined by $S = \mathbb{1}_{R>C}$ where C is the admissible rotation in degree. In our case, $C = 7.5°$. This value of admissible rotation is the 98%-level quantile from a sample of 1000 mechanical simulations, it is consistent with an industrial case situation when the failure is a rare event. The dataset of synthetic ground motions are filtered by a fictitious linear single-mode building at 5 Hz and damped at 2%. The fragility curve of the piping system is here expressed as a function of the pseudo-spectral acceleration of the initial set of the synthetic signals (i.e not filtered signals), calculated at 5 Hz and 1% damping ratio. The coarse estimation $\tilde{\theta}_0$ is estimated using 2000 computations with the linear FE model (an approximation of the costly nonlinear mechanical simulation). Figure 1 shows the results of a numerical benchmark consisting of 50 samples with 100 queries of a nonlinear mechanical computer code using AIS with a defensive parameter $\varepsilon = 10^{-3}$, the AIS procedure is initialized by sampling 20 signals with the density $q_{\tilde{\theta}_0,\varepsilon}$. AIS is compared to a naive algorithm consisting in 120 signals chosen at random in the unlabeled pool. Test error is evaluated on a dataset of 2000 computations of the mechanical response, sampled with the initial distribution p. As seen in Figure 1, the variance of the test error is significantly improved using AIS.

5 Conclusion

In this paper we introduce an active learning strategy in order to tackle the computational burden of mechanical simulations for fragility curve estimation. The main idea is to consider the statistical learning problem as a Monte Carlo estimation of the true expected risk from a data sample and to select the seism signals where to compute the mechanical response with an instrumental density that will end up reducing the variance of the training loss. This strategy was bench-marked on a FEM simulation of a safety water supply piping system of a French Pressurized Water Reactor, where performance between selecting seisms at random in an unlabeled dataset and AIS has been assessed.

References

Wei Chu, Martin Zinkevich, Lihong Li, Achint Thomas, and Belle Tseng. Unbiased online active learning in data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 195–203, New York, NY, USA, 2011. Association for Computing Machinery.

P. Hall, C.C. Heyde, Z.W. Birnbaum, and E. Lukacs. *Martingale Limit Theory and Its Application*. Communication and Behavior. Elsevier Science, 2014.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

F. Touboul, P. Sollogoub, and N. Blay. Seismic behaviour of piping systems with and without defects: experimental and numerical evaluations. *Nuclear Engineering and Design*, 192(2):243 – 260, 1999.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Procédures de tests multiples minimax pour la localisation d'une rupture dans un processus de Poisson

Fabrice Grela[1,2] & Magalie Fromont[1,3] & Ronan Le Guével[1,4]

[1] *Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*

[2] *fabrice.grela@univ-rennes2.fr*

[3] *magalie.fromont@univ-rennes2.fr*

[4] *ronan.leguevel@univ-rennes2.fr*

Résumé. On s'intéresse ici à la question de la localisation d'une rupture dans la loi d'un processus de Poisson observé sur l'intervalle $[0,1]$ et à la construction de procédures permettant d'estimer l'instant de rupture. On suppose que la rupture dans la loi du processus est caractérisée par un saut dans son intensité définie par rapport à une mesure Ldt où L est un entier strictement positif et dt est la mesure de Lebesgue sur $[0,1]$. Nous formulons ce problème de localisation comme un problème de tests multiples et présentons une étude minimax non asymptotique. En considérant la distance usuelle de $\mathbb{L}^2([0,1])$, nous établissons les vitesses de séparation minimax par famille, donnant un critère d'optimalité relatif à l'erreur de seconde espèce pour les tests multiples dont le FWER est contrôlé, sur différentes classes d'alternatives définies selon la connaissance ou non de la taille du saut. En supposant l'intensité de référence connue a priori, nous avons obtenu une vitesse de séparation minimax par famille d'ordre $L^{-1/2}$ dans le cas où la hauteur du saut est fixée et établi que l'adaptation en la hauteur du saut dégrade la vitesse d'un facteur $(\log\log L)^{1/2}$.

Mots-clés. Détection de rupture, processus de Poisson, tests multiples minimax, tests multiples adaptatifs.

Abstract. We here focus on the question of localizing a single change point in the distribution of a Poisson process observed on the interval $[0,1]$ and on the construction of change point estimation procedures. We assume that the abrupt change in the distribution of the process is characterized by a jump in its intensity defined with respect to some measure Ldt where L is a positive integer and dt is the Lebesgue measure over $[0,1]$. We formulate this change point localization problem as a multiple testing problem, and we present a nonasymptotic minimax study. By considering the usual distance of $\mathbb{L}^2([0,1])$, we establish the minimax family-wise separation rate, which provide a second kind error-related optimality criteria for multiple testing whose FWER is controlled, over various classes of alternatives defined according to whether the jump size is known or not. Assuming that the baseline intensity is fixed a priori, we prove that the knowledge of the jump size allows to obtain a minimax family-wise separation rate of order $L^{-1/2}$, and that the adaptation to the jump size deteriorates the rate by a $(\log\log L)^{-1/2}$ factor.

Keywords. Change point detection, Poisson process, minimax multiple tests, adaptive multiple tests.

1 Introduction

Les suites d'occurrences d'événements aléatoires observées sur un intervalle de temps sont généralement modélisées par des processus ponctuels. Il semble par exemple désormais admis que les occurrences de certaines tentatives d'intrusion dans des systèmes informatiques et que le trafic d'un réseau peuvent être modélisés par des processus de Poisson (c.f. Baldwin et al. (2017), Soltani et al. (2017)). La question de la détection de ruptures dans la loi de processus ponctuels peut donc traduire des problèmes concrets de détection de changements de régimes d'attaque dans les tentatives d'intrusion ou des comportements suspects (communications secrètes, virus) qui sont des enjeux-clés de la sécurité numérique.

Dans ce travail, nous nous intéressons plus particulièrement à la question de la localisation d'une rupture dans la loi d'un processus de Poisson. Chercher à estimer l'instant de saut une fois que la rupture a été détectée peut être vu comme la deuxième étape d'un problème de détection de rupture formulé dans un cadre hors-ligne.

On considère un processus de Poisson $N = (N_t)_{t\in[0,1]}$ observé sur l'intervalle $[0, 1]$, dont l'intensité λ est définie par rapport à une mesure Λ sur $[0, 1]$ et dont la loi est notée P_λ. On suppose que la mesure Λ vérifie $d\Lambda(t) = Ldt$, où $L \geq 1$ est un entier fixé et dt est la mesure de Lebesgue. Pour tout $a, b \in [0, 1]$, le nombre de points du processus N observé sur l'intervalle $(a, b]$ sera noté $N(a, b]$. On suppose que l'intensité λ appartient à l'ensemble $\mathcal{S}$ des fonctions de la forme $\lambda = \lambda_0 + \delta\mathbb{1}_{(\tau,1]}$ avec $\lambda_0 > 0$, $\delta \in (-\lambda_0, +\infty) \setminus \{0\}$ et $\tau \in (0, 1]$ ou $\delta \in (-\lambda_0, +\infty)$ et $\tau \in (0, 1)$. Dans toute la suite, on supposera également que l'intensité de référence λ_0 est fixée et est un paramètre connu du problème.

Selon que la taille du saut δ est connue ou non, les problèmes d'estimation de l'instant de rupture τ sont formulés ici comme des problèmes de tests multiples. Suivant la terminologie de Goeman and Solari (2010), on considère pour un entier non nul M une collection d'hypothèses $\mathcal{H}$ définie par $\mathcal{H} = \{H_k,\ k \in \{1, \dots, M\}\}$, où pour tout $k \in \{1, \dots, M\}$, H_k est inclus dans l'ensemble des fonctions constantes par morceaux avec au plus une discontinuité. On appelle ensemble des hypothèses vraies l'ensemble $\mathcal{T}(\lambda) = \{k \in \{1, \dots, M\},\ \lambda \in H_k\}$ et ensemble des hypothèses fausses, $\mathcal{F}(\lambda) = \{1, \dots, M\} \setminus \mathcal{T}(\lambda)$. Une procédure de tests multiples $\mathcal{R}$ associée à une collection d'hypothèses $\mathcal{H}$ est une statistique définie par un ensemble d'hypothèses simples rejetées $\mathcal{R} \subset \mathcal{H}$, dont l'objectif est d'inférer l'ensemble $\mathcal{F}(\lambda)$. Pour l'erreur de première espèce, on considère le critère du *Family-Wise Error Rate* (FWER) défini par

$$\mathrm{FWER}(\mathcal{R}, \mathcal{S}) = \sup_{\lambda\in\mathcal{S}} P_\lambda(\mathcal{R} \cap \mathcal{T}(\lambda) \neq \varnothing).$$

La vitesse de séparation minimax par famille ou *minimax Family-Wise Separation Rate* (mFWSR), introduite par Fromont, Lerasle et Reynaud-Bouret (2015), est un critère lié à l'erreur de seconde espèce à la base d'une théorie minimax pour les tests multiples dont le FWER est contrôlé. Pour la distance d_2 associée à la norme usuelle $\|\cdot\|_2$ de $\mathbb{L}^2([0,1])$ et $r>0$, on définit pour tout $\lambda \in \mathcal{S}$, $\mathcal{F}_r(\lambda) = \{H_k \in \mathcal{H} :\ d_2(\lambda, H_k) \geq r\}$. Pour $(\alpha,\beta) \in (0,1)^2$, une classe d'intensités $\mathcal{S}' \subset \mathcal{S}$ et une procédure de tests multiples $\mathcal{R}$ dont le FWER est contrôlé par α, la β-vitesse de séparation par famille de $\mathcal{R}$ sur $\mathcal{S}'$ est définie par

$$\mathrm{FWSR}_\beta(\mathcal{R}, \mathcal{S}') = \inf\{r>0 :\ \inf_{\lambda \in \mathcal{S}'} P_\lambda(\mathcal{F}_r(\lambda) \subset \mathcal{R}) \geq 1-\beta\}.$$

La $(\alpha,\beta)-$vitesse de séparation par famille minimax sur $\mathcal{S}'$ correspondante est alors définie par

$$\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}') = \inf_{\mathcal{R}:\ \mathrm{FWER}(\mathcal{R},\mathcal{S}) \leq \alpha} \mathrm{FWSR}_\beta(\mathcal{R}, \mathcal{S}').$$

On rappelle aussi un lemme prouvé dans l'article de Fromont, Lerasle et Reynaud-Bouret (2015) qui établit un lien entre le critère minimax pour les procédures de tests multiples et le critère minimax pour les procédures de test simple. Le critère minimax pour les tests simples, la vitesse de séparation minimax ou *minimax Separation Rate* (mSR), défini par Baraud (2002) traduit dans un cadre non asymptotique le critère minimax asymptotique pour les tests simples introduit par Ingster (1993).

Lemme 1. *Si la collection d'hypothèses $\mathcal{H}$ est fermée (c'est-à-dire que pour tout $H_k, H_{k'} \in \mathcal{H}$, on a $H_k \cap H_{k'} \in \mathcal{H}$), on obtient en notant $\cap\mathcal{H} = \cap_{k=1}^M H_k$,*

$$\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}') \geq \mathrm{mSR}^{\cap\mathcal{H}}_{\alpha,\beta}(\mathcal{S}') = \inf_{\phi_\alpha} \inf\left\{ r \geq 0 :\ \sup_{\lambda \in \mathcal{S}',\ d(\lambda, \cap\mathcal{H}) \geq r} P_\lambda(\phi_\alpha = 0) \leq \beta \right\},$$

le premier infimum étant pris sur tous les tests simples ϕ_α de niveau α de l'hypothèse nulle (H_0) "$\lambda \in \cap\mathcal{H}$", c'est-à-dire tels que $\sup_{\lambda \in \cap\mathcal{H}} P_\lambda(\phi_\alpha = 1) \leq \alpha$.

2 Localisation d'une rupture dans un processus de Poisson lorsque la taille du saut est connue

Dans cette partie, l'objectif est de construire une procédure de tests multiples pour localiser l'instant de rupture dans la loi d'un processus de Poisson lorsque la taille du saut est connue. Pour $\lambda_0 > 0$ et $\delta^* \in (-\lambda_0, +\infty) \setminus \{0\}$, on définit l'espace

$$\mathcal{S}[\lambda_0, \delta^*] = \{\lambda : [0,1] \to (0,+\infty),\ \exists \tau \in (0,1),\ \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau,1]}(t)\},$$

des intensités avec un saut de taille δ^* (connue) par rapport à λ_0 au temps τ.

Afin d'estimer l'instant de saut, on considère pour un entier $M \geq 1$ la collection d'hypothèses $\mathcal{H}_1 = \{H_k[\lambda_0, \delta^*],\ k \in \{1, \ldots, M\}\}$ où pour tout $k \in \{1, \ldots, M\}$,

$$H_k[\lambda_0, \delta^*] = \{\lambda :\ \exists \tau \in [k/M, 1],\ \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau,1]}(t)\}.$$

On remarque d'abord que la collection $\mathcal{H}_1$ est fermée car les hypothèses sont emboîtées:

$$H_M[\lambda_0, \delta^*] \subset H_{M-1}[\lambda_0, \delta^*] \subset \ldots \subset H_1[\lambda_0, \delta^*].$$

En particulier, $\cap\mathcal{H}_1 = H_M[\lambda_0, \delta^*] = \{\lambda_0\}$ et d'après le Lemme 1, $\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}[\lambda_0, \delta^*]) \geq \mathrm{mSR}^{\{\lambda_0\}}_{\alpha,\beta}(\mathcal{S}[\lambda_0, \delta^*])$. L'étude du problème de détection d'une rupture dans la loi d'un processus de Poisson sous l'angle des tests minimax non asymptotiques menée dans Fromont, Grela et Le Guével (2021) permet d'obtenir la proposition suivante.

Proposition 2 (Borne inférieure minimax)**.** *Soit* $(\alpha, \beta) \in (0,1)^2$ *tel que* $\alpha+\beta < 1$, $\lambda_0 > 0$ *et* $\delta^* \in (-\lambda_0, +\infty) \setminus \{0\}$. *Pour tout* $L \geq \lambda_0 \log C_{\alpha,\beta}/\delta^{*2}$,

$$\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}[\lambda_0, \delta^*]) \geq \sqrt{\frac{\lambda_0 \log C_{\alpha,\beta}}{L}} \quad \textit{où} \quad C_{\alpha,\beta} = 1 + 4(1-\alpha-\beta)^2.$$

Pour montrer que cette borne inférieure est d'ordre optimal, nous construisons une procédure de tests multiples dont le FWER est contrôlé par α et dont le FWSR atteint (à une constante près) la borne inférieure ci-dessus. Pour cela, on considère pour tout $k \in \{1, \ldots, M\}$ le problème de test simple pour l'hypothèse nulle $H_k[\lambda_0, \delta^*]$ contre l'alternative $\mathcal{S}[\lambda_0, \delta^*] \setminus H_k[\lambda_0, \delta^*]$. Ainsi, pour tout $k \in \{1, \ldots, M\}$, on définit $\phi_k^{(1)}$ par

$$\phi_k^{(1)} = \mathbb{1}_{D_k(\delta^*) > q_k(1-\alpha)},$$

où

$$D_k(\delta^*) = \sup_{t \in [0, k/M]} \left(\mathrm{sgn}(\delta^*) \left(N\left(t, \frac{k}{M}\right] - \lambda_0 L \left(\frac{k}{M} - t \right) \right) - \frac{|\delta^*|}{2} L \left(\frac{k}{M} - t \right) \right),$$

et $q_k(1-\alpha)$ est le $(1-\alpha)$-quantile de $D_k(\delta^*)$ sous $H_k[\lambda_0, \delta^*]$. Ces tests simples, qui prennent en compte la connaissance de la taille du saut δ^*, sont basés sur l'agrégation de statistiques de comptage driftées et s'inspirent des tests minimax pour la détection d'une rupture introduits par Fromont, Grela et Le Guével (2021). On considère $\hat{k} = \left(\sup\{k' \in \{1, \ldots, M\},\ \phi_{k'}^{(1)} = 0\} + 1\right) \wedge M$. La procédure de tests multiples $\mathcal{R}^{(1)}$ est alors définie par

$$\mathcal{R}^{(1)} = \{H_k[\lambda_0, \delta^*] :\ k \geq \hat{k}\}. \tag{1}$$

Le résultat ci-dessous utilise l'expression de la loi de probabilité et des inégalités exponentielles pour les suprema de processus de Poisson avec drift obtenues par Pykes (1959).

Théorème 3 (Borne supérieure minimax). *Soit* $(\alpha,\beta)\in(0,1)^2$, $\lambda_0>0$, $\delta^*\in(-\lambda_0,+\infty)\setminus\{0\}$ *et* $L\geq 1$. *Il existe des constantes positives* $L_0(\lambda_0,\delta^*,\alpha,\beta)$ *et* $C(\lambda_0,\delta^*,\alpha,\beta)$ *telles que pour tout* $L\geq L_0(\lambda_0,\delta^*,\alpha,\beta)$, *la procédure de tests multiples* $\mathcal{R}^{(1)}$ *définie par* (1) *satisfait*

$$\mathrm{FWER}(\mathcal{R}^{(1)},\mathcal{S})\leq\alpha,\ \ \textit{et}\ \ \mathrm{FWSR}_\beta(\mathcal{R}^{(1)},\mathcal{S}[\lambda_0,\delta^*])\leq\frac{C(\lambda_0,\delta^*,\alpha,\beta)}{\sqrt{L}}.$$

En particulier, $\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}[\lambda_0,\delta^*]))\leq C(\lambda_0,\delta^*,\alpha,\beta)/\sqrt{L}$.

3 Adaptation à la position et la taille du saut simultanément

Dans cette section, on s'intéresse à la question de l'adaptation en la position et la taille du saut. Pour $\lambda_0>0$ et $R>\lambda_0$, on introduit l'espace

$$\mathcal{S}[\lambda_0,R]=\{\lambda:\ \exists(\delta,\tau)\in((-\lambda_0,R-\lambda_0]\setminus\{0\})\times(0,1),\ \lambda(t)=\lambda_0+\delta\mathbb{1}_{(\tau,1]}(t)\},$$

des intensités bornées par R avec un saut de taille δ par rapport à λ_0 au temps τ. Afin d'estimer l'instant de saut, on considère pour un entier $M\geq 1$ la collection d'hypothèses $\mathcal{H}_2=\{H_k[\lambda_0,R],\ k\in\{1,\dots,M\}\}$ où pour tout $k\in\{1,\dots,M\}$,

$$H_k[\lambda_0,R]=\{\lambda:\ \exists(\delta,\tau)\in((-\lambda_0,R-\lambda_0]\setminus\{0\})\times[k/M,1],\ \lambda(t)=\lambda_0+\delta\mathbb{1}_{(\tau,1]}(t)\}.$$

La collection d'hypothèses $\mathcal{H}_2$ étant également fermée, le Lemme 1 et l'étude du problème de détection d'une rupture menée dans Fromont, Grela et Le Guével (2021) conduisent au résultat suivant.

Proposition 4 (Borne inférieure minimax). *Soit* $(\alpha,\beta)\in(0,1)^2$ *tel que* $\alpha+\beta<1/2$, $\lambda_0>0$ *et* $R>\lambda_0$. *Il existe* $L_0(\alpha,\beta,\lambda_0,R)>0$ *telle que pour tout* $L\geq L_0(\alpha,\beta,\lambda_0,R)$,

$$\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}[\lambda_0,R])\geq\sqrt{\frac{\lambda_0\log\log L}{L}}.$$

Soit $k\in\{1,\dots,M\}$. Comme dans la Section 2, on construit un test $\phi_k^{(2)}$ pour l'hypothèse nulle $H_k[\lambda_0,R]$ contre l'alternative $\mathcal{S}[\lambda_0,R]\setminus H_k[\lambda_0,R]$:

$$\phi_k^{(2)}=\mathbb{1}_{\max_{j\in\{1,\dots,\lfloor\log_2 L\rfloor\}}\left(T_{j,k}-t_{j,k}\left(1-\frac{\alpha}{\lfloor\log_2 L\rfloor}\right)\right)>0},$$

où pour tout $j\in\{1,\dots,\lfloor\log_2 L\rfloor\}$,

$$T_{j,k}=\frac{2^jM}{L^2k}\left(N\left(\frac{k}{M}\left(1-\frac{1}{2^j}\right),\frac{k}{M}\right]^2-N\left(\frac{k}{M}\left(1-\frac{1}{2^j}\right),\frac{k}{M}\right]\right)-\frac{2\lambda_0}{L}N\left(\frac{k}{M}\left(1-\frac{1}{2^j}\right),\frac{k}{M}\right]+\frac{\lambda_0^2k}{2^jM},$$

et $t_{j,k}(u)$ est le u-quantile de $T_{j,k}$ sous $H_k[\lambda_0, R]$.
On considère $\hat{k} = \left(\sup\{k' \in \{1, \dots, M\},\ \phi_{k'}^{(2)} = 0\} + 1\right) \wedge M$. La procédure de tests multiples $\mathcal{R}^{(2)}$ est alors définie par

$$\mathcal{R}^{(2)} = \{H_k[\lambda_0, R],\ k \geq \hat{k}\}. \tag{2}$$

La proposition suivante, basée sur de nouvelles inégalités exponentielles pour les modules d'oscillations de martingales et martingales carrées démontrées par Le Guével (2020), fournit une borne supérieure pour la vitesse de séparation minimax par famille. Avec la borne inférieure obtenue dans la Proposition 4, ces résultats mettent en évidence une perte inévitable de l'ordre d'un facteur $(\log\log L)^{1/2}$ dans la vitesse de séparation minimax par famille sur $\mathcal{S}[\lambda_0, R]$.

Théorème 5 (Borne supérieure minimax)**.** *Soit $(\alpha, \beta) \in (0,1)^2$, $\lambda_0 > 0$, $R > \lambda_0$ et soit $L \geq 3$. Il existe des constantes positives $L_0(\lambda_0, \alpha, \beta, R)$ et $C(\lambda_0, \alpha, \beta, R)$ telles que pour tout $L \geq L_0(\lambda_0, \alpha, \beta, R)$, la procédure de tests multiples $\mathcal{R}^{(2)}$ définie par (2) satisfait*

$$\mathrm{FWER}(\mathcal{R}^{(2)}, \mathcal{S}) \leq \alpha \quad \textit{et} \quad \mathrm{FWSR}_\beta(\mathcal{R}^{(2)}, \mathcal{S}[\lambda_0, R]) \leq C(\lambda_0, \alpha, \beta, R)\sqrt{\frac{\log\log L}{L}}.$$

En particulier, $\mathrm{mFWSR}_{\alpha,\beta}(\mathcal{S}[\lambda_0, R]) \leq C(\lambda_0, \alpha, \beta, R)\sqrt{(\log\log L)/L}$.

Note: Ce travail a été réalisé avec le soutien de la D.G.A. et de la région Bretagne.

Bibliographie

Baldwin, A., Gheyas, I., Ioannidis, C., Pym, D. and Williams, J. (2017), Contagion in cybersecurity attacks, *J. Oper. Res. Soc.*, 68(780).
Baraud, Y. (2002), Non-asymptotic minimax rates of testing in signal detection, *Bernoulli*, 8(5), 577-606.
Fromont, M., Grela, F. and Le Guével, R. (2021), Minimax and adaptive tests for detecting abrupt and possibly transitory changes in a Poisson process, *ArXiv*.
Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2015), Family-Wise Separation Rates for multiple testing, *Ann. Statist.*, 44(6), 2533-2563.
Goeman, J.J. and Solari, A. (2010), The sequential rejection principle of family wise error control, *Ann. Statist.*, 6, 3782–3810.
Ingster, Yu. I. (1993), Asymptotically minimax testing for nonparametric alternatives I-II-III. *Math. Meth. Stat.*, 2, 85–114, 171–189, 249–268.
Le Guével, R. (2020), Exponential inequalities for the supremum of some counting processes and their square martingales, *https://hal.archives-ouvertes.fr/hal-02275583*.
Pyke, R. (1959), The supremum and infimum of the Poisson process, *A. Math. Stat.*, 30.
Soltani, R., Goeckel, D., Towsley, D. and Houmansadr, A. (2017), Covert communications on poisson packet channels, *Institute of Electrical and Electronics Engineers*, pp. 1046-1052.

What does LIME really see in images?

Damien Garreau [1] & Dina Mardaoui [2]

[1] *Université Côte d'Azur, Inria, CNRS, LJAD, France*
Email: damien.garreau@unice.fr
[2] *Polytech Nice, France*
Email: dina.mardaoui@etu.univ-cotedazur.fr

Résumé. La capacité des machines à traiter des images a franchi un cap décisif ces dernières années. Ce bond en performance s'est accompagné d'une complexification des modèles utilisés qui rend très difficile la compréhension d'une prédiction individuelle. Nous analysons ici LIME, une méthode proposé en 2016 pour expliquer les prédictions d'un modèle sur des images, sans *a priori* sur ce modèle. En particulier, nous montrons que lorsque le nombre d'exemples perturbés est grand, les explications fournies par LIME convergent vers une explication limite dont nous dérivons une expression. Pour certains modèles simples (détecteur de forme et linéaire), cette expression est suffisamment simple pour nous permettre de montrer que LIME fournit des explications qui font réellement sens. Nous montrons également le lien avec une autre méthode d'interprétabilité, les gradients intégrés.

Mots-clés. Machine learning, interprétabilité, vision par ordinateur.

Abstract. The performance of modern algorithms on certain computer vision tasks such as object recognition is now close to that of humans. This success was achieved at the price of complicated architectures depending on millions of parameters and it has become quite challenging to understand how particular predictions are made. Interpretability methods propose to give us this understanding. In this paper, we study LIME, perhaps one of the most popular. On the theoretical side, we show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which we give an explicit expression. We further this study for elementary shape detectors and linear models. As a consequence of this analysis, we uncover a connection between LIME and integrated gradients, another interpretability method.

Keywords. Machine learning, interpretability, computer vision.

1 Introduction

In this paper, we study the image version of LIME [Local Interpretable Model-agnostic Explanations, Ribeiro et al., 2016]. Let us recall briefly how it operates: in order to explain the prediction of a model f for example ξ, LIME

1. decomposes ξ in d superpixels, that is, small homogeneous image patches;
2. creates a number of new images $x_1, \ldots, x_n$ by *randomly turning on and off* these superpixels (see Figure 1);
3. queries the model, getting predictions $y_i = f(x_i)$;
4. builds a local weighted surrogate model $\hat{\beta}_n$ fitting the y_is to the presence or absence of superpixels.

Each coefficient of $\hat{\beta}_n$ is associated to a superpixel of the original image ξ and, intuitively, the more positive the more important the superpixel is for the prediction at ξ according to LIME. Generally, the user visualizes $\hat{\beta}_n$ by highlighting the superpixels associated to the top positive coefficients.

The central question underlying this work is that of the soundness of LIME for explaining simple models: before using LIME on deep neural networks, are we sure that the explanations provided make sense for the most simple models? Can we guarantee it theoretically?

Related work. The present work follows the line of ideas initiated by Garreau and von Luxburg [2020a,b] for the tabular data version of LIME and later extended to text data by Mardaoui and Garreau [2021].

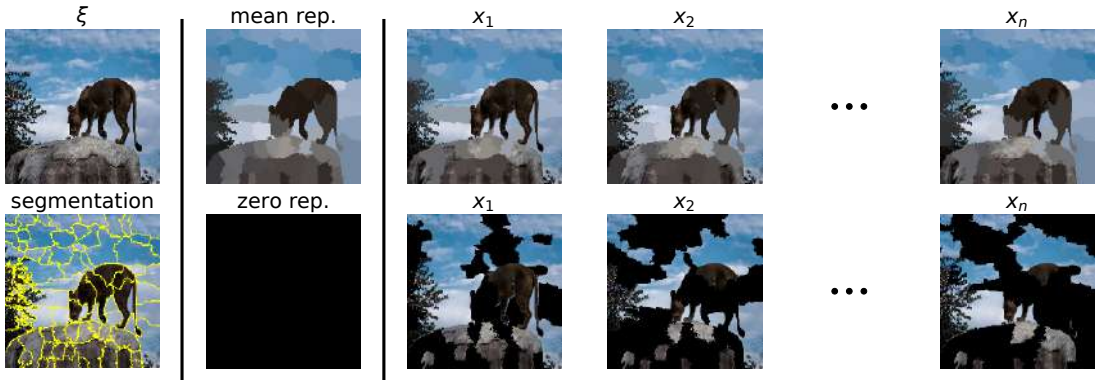


Figure 1: Sampling procedure of LIME for images. The image to explain, ξ, is first split into $d = 72$ superpixels (*lower left corner*). A replacement image $\overline{\xi}$ is computed, which is by default the mean of ξ on each superpixel (*top row*). This replacement image can also be filled uniformly with a pre-determined color (*bottom row:* replacement with the color black). Then, for each new example x_i with $1 \leq i \leq n$, the superpixels are randomly switched depending on the throw of d independent Bernoulli 1/2 random variables.

2 Main results

When the number of new samples n is large, we expect the empirical explanations provided by LIME to stabilize. Our first result formalizes this intuition.

Theorem 1 (Concentration of $\hat{\beta}_n$). *Assume that f is bounded by a constant $M > 0$ on $[0,1]^D$. Let $\epsilon > 0$ and $\eta \in (0,1)$. Let d be the number of superpixels. Then, there exists $\beta^f \in \mathbb{R}^{d+1}$ such that, for every*

$$n \gtrsim \max(M, M^2)\epsilon^{-2} d^7 \mathrm{e}^{\frac{4}{\nu^2}} \log \frac{8d}{\eta}\,,$$

we have $\mathbb{P}(\|\hat{\beta}_n - \beta^f\| \geq \epsilon) \leq \eta$.

We refer to the full version of this paper for a complete statement (we omitted numerical constants and the intercept for clarity) [Garreau and Mardaoui, 2021]. Intuitively, Theorem 1 means that when n is large, $\hat{\beta}_n$ stabilizes around β^f. Thus we can focus on β^f to study LIME. The main limitation of Theorem 1 is the dependency on d and ν: the control that we achieve on $\|\hat{\beta}_n - \beta^f\|$ is quite poor whenever d is too large or ν is too small. Note also that $\hat{\beta}_n$ is given by the *non-regularized* version of LIME.

It is possible to derive an explicit expression for β^f [Garreau and Mardaoui, 2021], but it is quite involved. Here, we only show how to compute β^f when f is linear. In that case, β^f takes a very simple form:

Proposition 1 (Computation of β^f, linear case). *Assume that $f(x) = \sum_{u=1}^{D} \lambda_u x_u$. Then, for any $1 \leq j \leq d$,*

$$\beta_j^f = \sum_{u \in J_j} \lambda_u \cdot (\xi_u - \overline{\xi}_u)\,,$$

where the J_j are the superpixels used by LIME.

When $\overline{\xi} = 0$, the coefficients take a very simple expression. Namely, the interpretable coefficient associated to superpixel J_j is **the sum of the coefficients of f multiplied by the pixel values on the superpixel**. If another replacement is chosen, then the *normalized* values of the pixel is taken into account in this product. This seems to make a lot of sense: let us say that the coefficients of f take large positive values on superpixel j. Then LIME will give a high interpretable coefficient to this superpixel, unless the pixel values are small (or very close to the replacement color if another replacement is chosen).

3 Approximated explanations

An interesting question in light of the results of the previous question is the following: if we approximate f by a linear function with coefficients λ between ξ and $\overline{\xi}$, are the explanations given by Proposition 1 close to the LIME explanations?

The most natural linear approximation of a function is given by its Taylor expansion truncated at order one. Since we want to approximate $f(x)$, where x is somewhere between ξ and $\overline{\xi}$, we could write, for instance, that $f(x) \approx f(\xi) + \nabla f(\xi)^\top (x - \xi)$. There are two main objection in doing so in the present case. First, we do not expect f to be linear between ξ and $\overline{\xi}$, and taking just one gradient would lead to a poor approximation.

Second, it is a well-known phenomenon in modern architecture that the gradient of the model with respect to the input can *saturate* when the network is confident in the prediction for certain activation functions. Since from our point of view f is a black-box model, we do not have information on the activation functions (in fact, we do not even assume that f is a neural network). Therefore gradients taken at ξ or $\overline{\xi}$ can be zero, giving us essentially no information on the behavior of f.

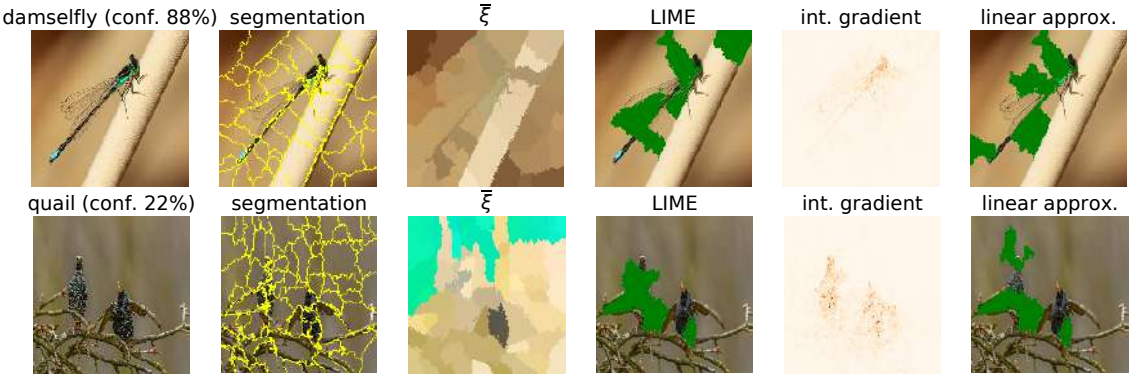


Figure 2: Comparing the explanations given by LIME *vs* approximate explanations obtained by summing the integrated gradient over the LIME superpixels. Here we explain the top predicted class for images of the ILSVRC2017 test data with the InceptionV3 network. In both cases, we showcase the top five positive coefficients. Qualitatively, the explanations obtained are quite similar, identifying close superpixels when they are not matching exactly.

For both these reasons, we build a linear approximation of f between ξ and $\overline{\xi}$ using the *averaged gradients on a linear path* between ξ and $\overline{\xi}$. Formally, we define

$$g_u := \int_0^1 \frac{\partial f((1-\alpha)\xi + \alpha\overline{\xi})}{\partial x_u} \, \mathrm{d}\alpha \tag{1}$$

the averaged gradient at pixel u. We approximate this integral by the Riemann sum g_u^{approx}. Subsequently, we approximate $f(x)$ by $(x - \overline{\xi})^\top g^{\text{approx}} + f(\overline{\xi})$. Applying Proposition 1 to this approximation we obtain the approximate explanations

$$\forall 1 \leq j \leq d, \quad \beta_j^{\text{approx}} = \sum_{u \in J_j} (\xi_u - \overline{\xi}_u) \cdot g_u^{\text{approx}} \, . \tag{2}$$

Inside the sum, we recognize the definition of *integrated gradients* between ξ and $\overline{\xi}$ [Sundararajan et al., 2017], another interpretability method. Eq. (2) therefore corresponds to **the sum of integrated gradients over superpixel** j.

Without being a perfect match, we observe a **substantial overlap between the LIME explanations and the approximated explanations** for all the models and datasets that we tried. This is particularly striking for simple models. More precisely, the Jaccard similarities observed are several times higher than what a random guess would produce. This is surprising since we are considering a linear approximation of highly non-linear functions. As a matter of fact, the exact values of the interpretable coefficients are quite different. Nevertheless, they are sufficiently close so that the sets of superpixels identified by both methods are consistently overlapping.

We notice that this link seems to weaken when the models become too complex, while still a third of identified superpixels are common for InceptionV3. However, visual inspection reveals that the superpixels identified by both methods remain close from each other even when they are distinct (see Figure 2).

References

A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

D. Garreau and D. Mardaoui. What does LIME really see in images? *arXiv preprint arXiv:2102.06307*, 2021.

D. Garreau and U. von Luxburg. Explaining the explainer: A first theoretical analysis of LIME. In *Proceedings of the 33rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1287–1296, 2020a.

D. Garreau and U. von Luxburg. Looking Deeper into Tabular LIME. *arXiv preprint arXiv:2008.11092*, 2020b.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

D. Mardaoui and D. Garreau. An Analysis of LIME for Text Data. In *AISTATS (to appear)*, 2021.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR*, pages 4510–4520, 2018.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Test d'hypothèse complexe appliqué à l'analyse de l'expression différentielle pour données RNA-seq en cellule unique

Marine Gauthier[1,2,3], Denis Agniel[5,6],
Rodolphe Thiébaut[1,2,3,4], Boris P. Hejblum[1,2,3]

[1] *Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, Bordeaux, France,* [2] *INRIA Bordeaux Sud Ouest, Talence, France* [3] *Vaccine Research Institute, Créteil, France* [4] *CHU, Bordeaux, France* [5] *Rand Corporation, Santa Monica (CA), USA* [6] *Harvard Medical School, Boston (MA), USA*
marine.gauthier@u-bordeaux.fr, denis.agniel@gmail.com, rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr,

Résumé. Le séquençage d'ARN en cellule unique (*scRNA-seq*) mesure l'expression des gènes à l'échelle unicellulaire. Les méthodes d'analyse d'expression différentielle (DEA) pour données RNA-seq en cellule unique reposent souvent sur des hypothèses distributionnelles difficiles à vérifier en pratique. Alors que la complexité grandissante des études cliniques et biologiques exige une plus grande flexibilité des approches, la majorité des méthodes existantes abordent seulement la comparaison entre deux conditions. Nous proposons donc une nouvelle approche, appelée `ccdf` permettant de tester l'association de l'expression d'un gène, quelque soit la distribution de cette dernière, avec une ou plusieurs variables explicatives d'intérêt (continues ou discrètes), potentiellement ajustées sur des covariables. Pour tester ces hypothèses complexes, `ccdf` utilise un test d'indépendance conditionnelle s'appuyant sur la fonction de répartition conditionnelle, estimée à l'aide de régressions multiples. `ccdf` comprend un test asymptotique ainsi qu'un test par permutations (lorsque le nombre de cellules observées n'est pas suffisamment grand). `ccdf` présente de bonnes performances statistiques dans divers scénarios d'analyse, incluant des plans d'expériences complexes (*i.e.* au-delà de la simple comparaison entre deux conditions) – tout en conservant des performances compétitives avec l'état de l'art lorsque l'on teste deux conditions.

Mots-clés. single-cell ; fonction de répartition conditionnelle ; test d'indépendance conditionnelle ; Analyse de l'expression différentielle ; test asymptotique

Abstract. Single-cell RNA-seq (scRNA-seq) quantifies gene expression at the cell resolution. State-of-the-art methods for scRNA-seq Differential Expression Analysis (DEA) often rely on strong distributional assumptions that are difficult to verify in practice. Furthermore, while the increasing complexity of clinical and biological single-cell studies calls for greater tool versatility, the majority of existing methods only tackles the comparison between two conditions. We propose a novel, distribution-free, and flexible approach to DEA for single-cell RNA-seq data. This new method, called `ccdf`, tests the

association of each gene expression with one or many variables of interest (that can be either continuous or discrete), while potentially adjusting on additional covariates. To test such complex hypotheses, `ccdf` uses a conditional independence test relying on the conditional cumulative distribution function, estimated through multiple regressions. `ccdf` includes an asymptotic test as well as a permutation test (when the number of observed cell is not sufficiently large). `ccdf` exhibits good statistical performance in various simulation scenarios considering complex experimental designs (*i.e.* beyond the two condition comparison), while retaining competitive performance with the state-of-the-art in a two condition benchmark.

Keywords. single-cell; conditional cumulative distribution function; conditional independence test; differential expression analysis; asymptotic test

1 Limites des méthodes scRNA-seq existantes

Le séquençage de l'ARN en cellule unique (*scRNA-seq*) permet de mesurer simultanément les niveaux d'expression des gènes à la résolution de centaines voire de milliers de cellules individuelles, contrairement au séquençage de l'ARN en masse (*bulk RNA-seq*) qui mesure l'expression moyenne d'un ensemble de cellules. Notre objectif est de réaliser une analyse différentielle de l'expression génique (DEA), c'est-à-dire trouver quels gènes sont différentiellement exprimés en fonction de certaines variables d'intérêt. Un gène est dit différentiellement exprimé (DE) si son expression est significativement associée aux variations d'un facteur d'intérêt. En plus d'une grande proportion de zéros observés, appelés "dropouts" [1], les données scRNA-seq peuvent générer des distributions multimodales et hétérogènes. Par conséquent, caractériser la distribution de l'expression génique par cellule est délicat. D'une part, plusieurs méthodes paramétriques ont été proposées, comme MAST [2], SCDE [3], scDD [4] et DEsingle [5]. D'autre part, des approches non-paramétriques ont été développées pour mieux modéliser les distributions, comme EMDomics [6] et plus récemment SigEMD [7]. Reposant sur de fortes hypothèses distributionnelles difficiles à vérifier en pratique, les outils paramétriques se heurtent à des problèmes méthodologiques. Enfin, les méthodes précédemment citées, qu'elles soient paramétriques ou non-paramétriques, ne sont pas capables d'analyser des designs expérimentaux complexes, ce qui les rend très restrictives dans leur utilisation.

2 Motivation

Un nombre important d'expériences biologiques générant des données scRNA-seq reste dans le cadre classique de l'analyse différentielle consistant à l'étude de 2 conditions (par exemple, vaccin/placebo). Il est pourtant envisageable de vouloir tester l'association de

l'expression d'un gène avec une variable discrète à plus de 2 classes (sous-populations cellulaires ou différentes doses vaccinales par exemple), ou bien encore de vouloir tester l'association avec une variable continue, par exemple avec l'expression d'un gène spécifique ou d'autres biomarqueurs identifiés. Le jeu de données réelles qui a motivé notre travail est composé de 18 513 gènes mesurés dans 2 914 cellules dendritiques (DC). Quatre sous-populations ont au préalable été annotées par le Vaccine Research Institute : DC1, DC2 & DC3, pDC et preDC. On cherche à identifier quels gènes s'expriment différentiellement dans les 4 sous-populations cellulaires. Les méthodes actuelles ne permettant pas de prendre en compte un tel schéma d'étude, nous proposons ici une nouvelle méthode d'analyse différentielle pour données scRNA-seq. Cette dernière ne nécessite aucune hypothèse distributionnelle sur les données d'expression (distribution débatue au sein de la communauté scientifique [8]), et peut tester l'association de l'expression génique avec une ou plusieurs variables d'intérêt, qu'elles soient continues et/ou discrètes, en ajustant sur de potentielles covariables (par exemple le type cellulaire). Notre méthode se voudra être la plus flexible possible pour s'adapter plus grand nombre de plans expérimentaux possible.

3 Méthode

Tester l'association de la variable aléatoire Y, représentant l'expression d'un gène, avec un facteur ou un groupe de facteurs d'intérêt X connu, soit discret (*e.g.* comparaisons multiples) soit continu, étant données des covariables Z également connu, équivaut à tester l'indépendance conditionnelle entre Y et X sachant Z :

$$H_0 : Y \perp X \mid Z \tag{1}$$

Le test d'indépendance conditionnelle que nous proposons repose sur des fonctions de répartition conditionnelles. En effet, si un groupe de facteurs est associé à l'expression du gène, la conséquence immédiate est que la fonction de répartition conditionnelle de l'expression du gène est différente de la fonction de répartition marginale, c'est-à-dire sans conditionnement par le groupe de facteurs. Ainsi, l'hypothèse nulle peut s'écrire de la façon suivante :

$$H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z) \tag{2}$$

où la fonction de répartition conditionnelle de Y sachant X et Z est définie comme $F_{Y|X,Z}(y, x, z) = \mathbb{P}(Y \leq y \mid X = x, Z = z)$. S'il n'y a pas de covariables, le test d'indépendance conditionnelle se transforme en un simple test d'indépendance puisque nous testons alors l'hypothèse nulle $Y \perp X$ qui est équivalente au test $F_{Y|X}(y, x) = F_Y(y)$.

3.1 Statistique de test

Notons $\boldsymbol{Y}^g = (Y_1^g, ..., Y_n^g)$ un vecteur contenant un compte normalisé (i.e. expression génique) pour le gène g dans les n cellules et $\boldsymbol{X}^g = (\boldsymbol{X}_1^g, ..., \boldsymbol{X}_n^g)$ une matrice $s \times n$

codant la ou les conditions à tester (qui peuvent être continues ou discrètes). On peut vouloir ajouter des variables exogènes, qui n'ont pas à être testées mais sur lesquelles il est nécessaire d'ajuster le modèle. Soit $\boldsymbol{Z}^g = (\boldsymbol{Z}_1^g, ..., \boldsymbol{Z}_n^g))$ une $r \times n$ matrice des covariables continues ou discrètes à prendre en compte. Par souci de simplicité, nous omettons à présent la notation g car nous nous référons à la DEA au niveau du gène.

On pose $\boldsymbol{Y} \in [Y_{min}, Y_{max}]$ avec $Y_{min} = min(Y_1, ..., Y_n)$ et $Y_{max} = max(Y_1, ..., Y_n)$ alors $\omega_1, ..., \omega_p$ est une séquence ordonnée et régulière de p seuils entre Y_{min} et Y_{max}. Pour chaque ω_j avec $j = 1, ..., p$, la fonction de répartition conditionnelle $F_{Y|X,Z}(\omega_j \mid x, z)$ est égale à $\mathbb{P}(Y \leq \omega_j \mid X = x, Z = z)$ qui peut être réécrite comme une espérance conditionnelle : $F_{Y|X,Z}(\omega_j \mid x, z) = \mathbb{E}\left(\mathbb{1}_{\{Y \leq \omega_j\}} | X = x, Z = z\right)$ où $\mathbb{1}_{\{Y \leq \omega_j\}}$ est une variable aléatoire binaire qui est égale 1 si $Y \leq \omega_j$ et 0 sinon. Ensuite, pour une séquence de ω_js, on peut l'estimer par p régressions linéaires :

$$\mathbb{E}\left(\mathbb{1}_{\{y_i \leq \omega_j\}} | X = x_i, Z = z_i\right) = \beta_{0j} + \boldsymbol{\beta}_{1j}\boldsymbol{x}_i + \boldsymbol{\beta}_{2j}\boldsymbol{z}_i, \quad \forall i = 1, ..., n \tag{3}$$

Si X n'a pas de lien avec Y sachant Z, alors $\boldsymbol{\beta}_{1j}$ devrait être nul. Notre but est alors de tester : $H_0 : \boldsymbol{\beta}_{1j} = \mathbf{0}$, $j = 1, ..., p$ où $\boldsymbol{\beta}_{1j}$ fait référence à la régression de $\mathbb{1}_{\{y_i \leq \omega_j\}}$ sur $\boldsymbol{x}_i$, pour les seuils fixes $\omega_1, \omega_2, ..., \omega_p$. Enfin, nous transformons la matrice $\boldsymbol{\beta}_1$ de taille $s \times p$ en un vecteur $\boldsymbol{\gamma}^1$ de taille ps en concaténant les s lignes de $\boldsymbol{\beta}^1$ les unes après les autres. Nous proposons d'utiliser la statistique de test suivante afin de tester cette hypothèse nulle : $D = n\sum_{j=1}^{ps}\gamma_{1j}^2$

3.2 Estimation de la distribution sous l'hypothèse nulle

Après estimation de $\boldsymbol{\beta}_1$, et donc de $\boldsymbol{\gamma}_1$, le théorème central limite multivarié nous permet de monter que :

$$\sqrt{n}\left(\widehat{\boldsymbol{\gamma}_1} - \boldsymbol{\gamma}_1{}^*\right) \longrightarrow N(0, \Sigma) \tag{4}$$

où $\boldsymbol{\gamma}_1^*$ est l'espérance de $\boldsymbol{h}_i\tilde{\boldsymbol{y}}_i$ (concaténé en un vecteur). Sous l'hypothèse nulle, $\boldsymbol{\gamma}_1{}^* = \mathbf{0}$. Σ est une matrice de variance-covariance symétrique semi-définie positive de taille $ps \times ps$ qui peut être estimée par la méthode des moments : $\Sigma = n^{-1}\sum_{i=1}^{n}\mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{y}}_i) = E\{\mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{y}}_i)\}$. On peut ainsi montrer que la distribution asymptotique de $\widehat{D}_n$ sous l'hypothèse nulle est alors :

$$\widehat{D}_n \underset{n\to+\infty}{\longrightarrow} \sum_{j=1}^{ps}\widehat{a}_j\chi_1^2 \tag{5}$$

où $\widehat{a}_j$ est la j-ème valeur propre de la matrice de variance-covariance Σ.

La p-valeur est obtenue en comparant la statistique de test observée $\widehat{D}_n$ à la distribution de $\sum_{j=1}^{ps}\widehat{a}_j\chi_1^2$. En pratique, nous utilisons l'approximation de Davies [9] pour estimer le mélange de χ^2s, implémenté dans le package R `CompQuadForm` [10]. Il est à noter que

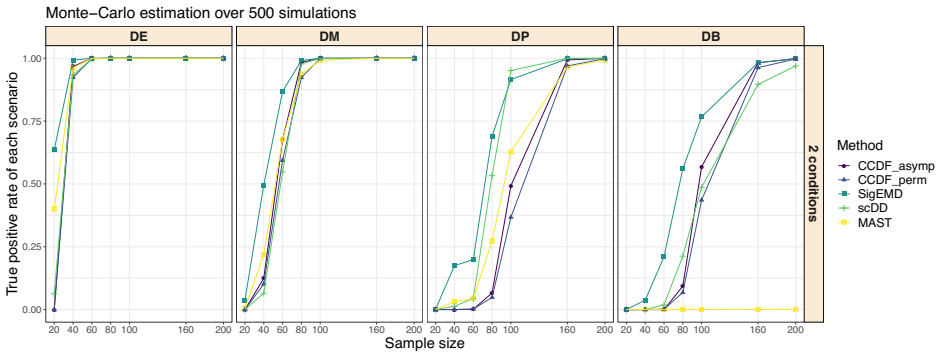


FIGURE 1 – Taux de vrais positifs pour chaque scénario (DE, DM, DP et DB)

nous obtenons une distribution asymptotique simple sans nous appuyer sur une quelconque hypothèse distributionnelle des données. Enfin, nous appliquons la correction de Benjamini-Hochberg [11].

Nous avons également développé un test par permutation pour le cas où le nombre de cellules observé est trop faible pour que le théorème central limite entre en action.

4 Resultats

Nous évaluons les performances de notre méthode `ccdf` en comparaison avec trois méthodes : `MAST`, `scDD`, et `SigEMD`, dans le cadre de la comparaison classique entre deux conditions. Pour cela nous générons des comptes à partir de distributions binomiales négatives et de mélanges de distributions binomiales négatives. 500 jeux de données sont simulées, composés de 10 000 gènes chacun, dont 1 000 sont exprimés différentiellement entre deux conditions, pour 7 tailles d'échantillon différentes allant de 20 à 200. Les observations sont réparties de manière égale entre les deux groupes. Nous nous sommes inspirés des quatre scénarios décrits par Korthauer et *et al.* [4] et avons donc simulé 250 gènes différentiellement distribués en moyenne (DE), 250 gènes différentiellement distribués en mode (DM), 250 gènes différentiellement distribués en proportion (DP) et 250 gènes différentiellement distribués en moyenne et en mode (DB). D'après la Figure 1, `ccdf`, `SigEMD` et `scDD` ont des performances comparables dans les quatre scénarios tandis que `MAST` ne parvient pas à détecter les gènes DB. Nous discuterons des différences de puissance statistique.

Ensuite, nous présenterons les résultats des performances de ces 4 méthodes lorsque nous testons l'association entre Y et X (2 conditions) conditionnellement à une variable continue. Nous illustrerons également les performances de `ccdf` lorsque nous devons tester plus de 2 conditions, les autres approches ne traitant pas ce cas particulier. Enfin, nous enrichirons cette étude de simulation par l'analyse de jeux de données de contrôle négatif et positif. Ces simulations plus complexes démontrent les capacités accrues de `ccdf` qui

conserve une puissance comparable à l'état de l'art tout en permettant d'analyser toute une panoplie de situations pour lesquelles il n'existe actuellement aucune alternative. La méthode décrite ici a été implémentée dans un package R appelé `ccdf`, disponible sur https ://github.com/Mgauth/ccdf et bientôt sur le CRAN.

Références

[1] Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome biology. 2020 ;21(1) :1–35.

[2] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST : a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology. 2015 ;16(1) :1–13.

[3] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods. 2014 ;11(7) :740–742.

[4] Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome biology. 2016 ;17(1) :222.

[5] Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018 ;34(18) :3223–3224.

[6] Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics : a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016 ;32(4) :533–541.

[7] Wang T, Nabavi S. SigEMD : A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018 ;145 :25–32.

[8] Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome biology. 2019 ;20(1) :1–16.

[9] Davies RB. The distribution of a linear combination of $\chi 2$ random variables. Journal of the Royal Statistical Society : Series C (Applied Statistics). 1980 ;29(3) :323–333.

[10] Duchesne P, De Micheaux PL. Computing the distribution of quadratic forms : Further comparisons between the Liu–Tang–Zhang approximation and exact methods. Computational Statistics & Data Analysis. 2010 ;54(4) :858–862.

[11] Benjamini Y, Hochberg Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal statistical society : series B (Methodological). 1995 ;57(1) :289–300.

K-MOM algorithme de clustering robuste

Edouard GENETAY [1] & Adrien SAUMARD [2] & Camille SAUMARD [3]

[1] *ENSAI, Campus de Ker-Lann,51 Rue Blaise Pascal BP 37203 à 35172 BRUZ Cedex, FRANCE et edouard.genetay@ensai.fr*

[3] *ENSAI, Campus de Ker-Lann,51 Rue Blaise Pascal BP 37203 à 35172 BRUZ Cedex, FRANCE et adrien.samard@ensai.fr*

[2] *Twice.AI®, 29 Avenue Jean-Janvier 35000 Rennes et csaumard@twice.ai*

Résumé. Les méthodes de clustering classiques telles que K-means souffrent d'un manque de robustesse à la présence de données aberrantes (outliers). Nous proposons un algorithme de clustering robuste basé sur l'utilisation de statistiques de type Median-Of-Means (MOM), une méthode qui s'est déjà avérée efficace en apprentissage supervisé robuste. L'implémentation de l'estimateur que nous proposons (K-MOM) consiste en la constitution de b sous-échantillons des observations, puis chaque sous-échantillon est clusterisé par l'algorithme de Lloyd classique initialisé par K-means++ pour finalement ne retenir que les centres de clusters qui ont réalisé la médiane de la distorsion K-means parmi les b sous-echantillons. Notre procédure a de meilleures performances que K-means et K-medoids en présence d'outliers ou de distributions à queue lourde. Dans ce contexte, K-MOM est comparable à K-medians et trimmed-K-means. Enfn, cette procédure peut aussi servir d'initialisation robuste à d'autres algorithmes.

Mots-clés. Clustering robuste, Perte empirique K-means, initialisation robuste, Breakdown point

Abstract. Classical clustering methods, such as K-means, suffer from a lack of robustness with respect to outliers. We propose a robust clustering algorithm based on Median-Of-Means statistics, a strategy that has been recently put to emphasis for efficient robust classification. The implementation of our estimator begins with the drawing of b subsamples of the observations, after that, each subsamples gets clustered by Lloyd's algorithm initialized by K-means++ and as final step the algorithm outputs the centroids that lead to the median value of the K-means loss among all subsample. Our procedure has better performances than K-means and K-medoids on corrupted or heavy-tailed data while being comparable to K-medians and trimmed-K-means. Moreover, our procedure supplies also a robust initialization method.

Keywords. Robust clustering, K-means loss, Benchmark, Breakdown point

1 K-MOM: algorithme de clustering robuste

La sensibilité de la plupart des approches de classification non-supervisée, telles que K-means ou modèles de mélange, à la qualité de l'initialisation d'une part et à la présence d'outliers d'autre part est bien connue. Nous proposons de ce fait d'associer l'estimateur MOM aux méthodes les plus répandues comme K-means puis de mesurer le gain apporté sur les performances de classification non-supervisée.

1.1 Median-of-Means : un estimateur robuste de la moyenne

Grâce aux travaux de Devroye Devroye et al. , on sait que Median-of-Means (MOM) est un estimateur de la moyenne μ qui se concentre optimalement autour μ.

Definition 1. Tout d'abord, on note $x_1^n := (x_1 \dots x_n) \in R^n$, S_n l'ensemble des permutation de $\{1, \dots, n\}$ muni de la mesure de probabilité uniforme μ, la médiane d'un ensemble $\mathcal{E} \subset R$, pour une mesure de probabilité λ sur $\mathcal{E}$, est notée med$\mathcal{E}$ et si cette médiane n'est pas unique, on prend la plus petite valeur médiane admissible :

$$\text{med}\mathcal{E} := \inf \left\{ t \in \mathcal{E} \,\middle|\, \lambda\left(\{y \in \mathcal{E} : y \leq t\}\right) \geq \frac{1}{2} \right\}$$

Soit b, t deux entiers tels que $bt = n$ et soit également $\bigcup_{k=1}^{b} B_k$ la partition de $\{1, \dots, n\}$ où $B_k := \{(k-1)t + i \,|1 \leq i \leq t\}$. Median-of-Means (MOM) est alors un estimateur réel randomisé défini de la façon suivante :

$$MOM_{n,b} : \begin{cases} R^n \times S_n & \longrightarrow & R \\ (x_1^n, \sigma) & \longmapsto & \text{med} \quad \left\{ \frac{1}{t} \sum_{i \in B_k} X_{\sigma(i)} : 1 \leq k \leq b \right\} \end{cases}$$

où en pratique, la permutation est un permutation aléatoire avec loi uniforme sur S_n.

Theorem 2. *Si $n > 5$ est un entier, $M > 0$, $\delta \in \left[2e^{-n/4}, 1/2\right)$, alors pour n'importe quelle variable aléatoire Y d'espérance μ et de variance $\text{Var}(Y) < M$, pour toute précision $\delta \geq e^{1-n/2}$, l'estimateur $MOM_{n,b}$ avec un nombre de blocs $b = \log(1/\delta)$ vérifie*

$$P\left(\left| MOM_{n,b}\left(Y_1^n\right) - \mu \right| > \sqrt{\frac{96M \log(1/\delta)}{n}} \right) \leq \delta$$

Or, d'après les travaux de Catoni, la meilleure concentration de la moyenne empirique est donnée par l'inégalité de Tchebychev si la seule hypothèse est $\text{Var}(Y) < M$. Cela fait de MOM un bien meilleur estimateur de la moyenne que la moyenne empirique. C'est pourquoi il paraît interessant de l'injecter dans les estimateurs tels que K-means, là où l'on a recours à une moyenne empirique (voir section 1.3). De plus, MOM octroie une plus grande robustesse aux outliers que la moyenne empirique d'après le survey de Mendelson et Lugosi.

1.2 Le bénéfice d'une version bootstrap de MOM

MOM recourt à une partition des données et clusteriser des données en K classes nécessite d'avoir suffisamment d'observations. Par conséquent, pour éviter une division excessive des données, nous utilisons une version bootstrap de MOM (voir équation 1). Par ailleurs, sur la base d'une définition du breakdown point proche de la définition du livre de Hampel et al :

Definition 3. Soit $\hat{\theta}_.$ un estimateur réel randomisé dont l'aléa est pris par rapport à l'espace probabilisé $(Q, \mathcal{Q}, Q)$ et dont la loi est invariante par permutation de ses arguments :

$$\forall x_1^n \in R^n, \forall \sigma \in S_n,\ \hat{\theta}_.(x_1 \dots x_n) \overset{\mathcal{D}}{=} \hat{\theta}_.(x_{\sigma(1)} \dots x_{\sigma(n)})$$

On appelle breakdown point probabiliste de $\hat{\theta}_.$, noté $\mathrm{BP}\left(\hat{\theta}_.\right)$, le plus petit nombre d'arguments de $\hat{\theta}_.$ à modifier de sorte que la déviation infligeable à l'estimateur soit infinie avec probabilité au moins $\frac{1}{2}$. On note également $x_1^n y_1^m := (x_1 \dots x_n, y_1, \dots, y_m)$,

$$\mathrm{BP}\left(\hat{\theta}_.\right) := \frac{1}{n} \inf \left\{ q \in N \,\middle|\, \exists x_1^n \in R^n, Q\left(\sup_{y_1^q \in R^q} |\hat{\theta}_.(x_1^{n-q} y_1^q) - \hat{\theta}_.(x_1^n)| = \infty \right) \geq \frac{1}{2} \right\}$$

Nous avons pu calculer la valeur exacte asymptotique du breakdown point d'une version bootstrap de MOM, définie ci-dessous :

Definition 4. Soit $x_1^n \in R^n$ un n-echantillon et notons $\left(X_j^*\right)_{1 \leq j \leq bt}$ $b \times t$ copies i.i.d. de la variable aléatoire qui vaut x_i avec probabilité $\frac{1}{n}$ pour tout $i \in \{1, \dots, n\}$, toutes définies sur l'espace probabilisé $(\Omega, \mathcal{T}, Q)$. Median-of-Means bootstrap ($MOMB$) est alors un estimateur réel randomisé défini comme suit :

$$MOMB_{n,b,t} : \left\{ \begin{array}{ccc} R^n \times \Omega & \longmapsto & R \\ (x_1^n, \omega) & \to & \mathrm{med} \ \left\{ \frac{1}{t} \sum_{j \in B_k} X_j^*(\omega) : 1 \leq k \leq b \right\} \end{array} \right. \tag{1}$$

où $B_k := \{(k-1)t + i \,|1 \leq i \leq t\}$

Proposition 5. *Pour $n, t, q \in N$, avec n le nombre de données, t la taille des blocs, q le nombre d'outliers, le breakdown point de $MOMB_{n,b,t}$ ne dépend que de la proportion d'outliers et de la taille des blocs puisqu'on peut choisir b de sorte à être le plus robuste possible du fait que* $\lim_{b \to \infty} \mathrm{BP}\left(MOMB_{n,b,t}\right) = 1 - \frac{1}{2^{1/t}} \underset{t \to \infty}{\sim} \frac{\log(2)}{t}$.

1.3 K-MOM: estimateur robuste de classfication non-supervisée

Nous proposons de remplacer la moyenne empirique qui apparaît dans le risque K-means : $\frac{1}{n} \sum_{i=1}^n \min_{1 \leq j \leq K} \|X_i - c_j\|_2^2$ par $MOMB_{n,b,t}$ comme suit :

$$\widehat{c}_{K,n,b,t} \in \underset{\mathbf{c}\in\mathcal{X}^K}{\operatorname{argmin}} \left[MOMB_{n,b,t} \left(\left(\min_{1\leq j\leq K} \|X_i - c_j\|_2^2 \right)_{1\leq i\leq n} \right) \right]$$

De plus, à l'instar de K-means, il est possible d'implémenter une procedure itérative telle que l'algorithme de Lloyd (voir algorithme (1)). Un bénéfice important de cette méthode est de fournir une procédure d'initialisation robuste en remplaçant l'algorithme de Lloyd par l'initialisation K-means++ uniquement, nous appelons cette initialisation K-MOM++.

Algorithm 1 K-MOM

Input: les données $\{X_1, \ldots, X_n\}$, b le nombre de blocs avec $t > K$ la taille des blocs

1. pour tout bloc j, $1 \leq j \leq b$:
   (a) Sélectionner au hasard uniformément et avec remise t observations
   (b) éxecuter l'algorithme de Lloyd initialisé par K-means++ pour obtenir K centres $\hat{\mathbf{c}}^{(j)}$
   (c) Calculer la perte K-means $\widehat{\mathrm{R_j}(\hat{\mathbf{c}})} = \frac{1}{n}\sum_{i=1}^{n} \min_{1\leq k\leq K} \left\| X_i - \hat{c}_k^{(j)} \right\|_2^2$ du bloc
2. Sélectionne le codeboob $\hat{\mathbf{c}}^{med} := \left(\widehat{c}_1^{(med)}, \ldots, \widehat{c}_K^{(med)}\right)$ du bloc dont la perte K-means est la valeur médiane parmi tous les blocs.

Output: $\hat{\mathbf{c}}^{med} := \left(\widehat{c}_1^{(med)}, \ldots, \widehat{c}_K^{(med)}\right)$

2 Simulations numériques

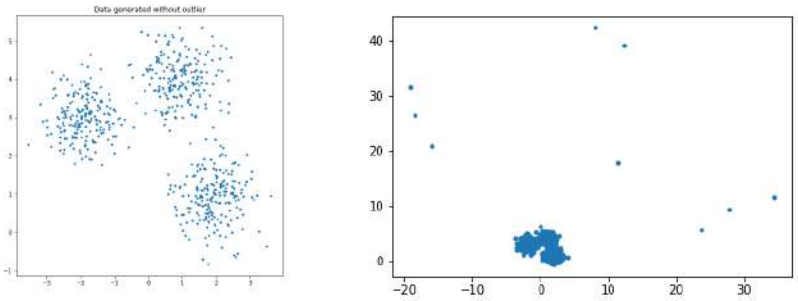
Pour mesurer l'efficacité des méthodes comparées, nous mesurons leurs performances d'estimation des centres des clusters grâce au RMSE (root mean square error) :

$$\mathrm{RMSE}\left(\hat{\mathbf{c}}, \mu\right) = \frac{1}{\sqrt{K}} \min_{\sigma\in S_K} \sqrt{\sum_{k=1}^{K} \left\| \hat{c}_{\sigma(k)} - \mu_k \right\|_2^2}$$

où $\hat{c}_k$ est le centres estimés de la classe k, pour rendre le RMSE insensible à la numérotation des classes, on prend le minimum parmi les $\sigma \in S_K$, l'ensemble des permutations d'ordre K et enfin μ_k est le centres théorique de la classe k.

Pour mesurer ces performances , les données sont simulées à partir de $K = 3$ gaussiennes multivariées de dimension $p = 2$. On tire dans chaque cluster $n_1 = n_2 = n_3 = 300$ points de variance $\sigma^2 = 0.6$ et de moyenne $\mu_1 = [1, 4]$, $\mu_2 = [2, 1]$ and $\mu_3 = [-2, 3]$. Parmi

ces $n = 900$ points, on choisit uniformément au hasard $n_{outlier}$ points dont on multiplie les coordonnées par $\beta > 0$. β contrôle la distance de ces outliers à leur cluster d'origine. Nous avons considéré 2 niveaux de corruption $n_{outlier} \in \{9, 27\}$ et 4 niveaux de distance $\beta \in \{5, 10, 20, 40\}$. La figure 1 illustre un jeu de données pour $n_{outlier} = 9$ et $\beta = 10$. Le nombre de clusters est supposé connu et $K = 3$.



(a) Données avant génération des outliers

(b) Données avec outliers

Figure 1: Illustration des données de simulation servant à comparer les algorithmes de clustering en présence d'outliers, dans ce cas $n_{outlier} = 9$ et $\beta = 10$

Algorithmes comparés : On compare notre algorithme K-MOM aux algorithmes suivants : K-means, trimmed-K-means, K-medians, K-medoids. Tous initialisés de trois façons : au hasard, par ROBIN et par l'initialisation K-MOM++. Les hyperparamètres de notre algorithme (b le nombre de blocs et t leur taille) sont pris tels que $t = 18$ et $b = 251$ de sortes que la probabilité que le bloc médian contienne un outliers soit inférieure à $0,05$.

3 Résultats empiriques

Les résultats présentés dans le tableau 1 ne montrent que les résultats dans le cas le plus défavorable ($n_{outlier} = 9$ et $\beta = 40$). Premièrement, on peut remarquer que l'initialisation K-MOM++ est bien meilleure que les procédures d'initialisation K-means++, l'initialisation au hasard et l'initialisation ROBIN. Deuxièmement, l'algorithme K-MOM est meilleur que K-medoids et trimmed-K-means et est comparable à K-medians.

Table 1: Performances des algorithmes de clustering et de leur initialisation en termes de RMSE dans le cas $n_{outlier} = 9$ et $\beta = 40$ La colonne P[G|init X] indique la proportion de résultats donnant un RMSE<2 pour chacun des algorithmes initialisé avec l'initialisation X. Cette façon de faire évite d'avoir des variances trop grandes et rendant les algorithmes incomparables.

	initialisé au hasard		initialisé avec ROBIN		initialisé avec K-MOM++	
algorithmes	P[G|random]	RMSE	P[G|ROBIN]	RMSE	P[G|K-MOM++]	RMSE
initialisation	0.2	0.75 (0.25)	0.25	0.48 (0.13)	**0.99**	0.37 (0.15)
K-means	0.0	0 (0)	0.0	0 (0)	0.0	0 (0)
K-medoids	0.45	1.04 (0.1)	0.45	1.03 (0.12)	0.56	1.06 (0.1)
trimmed-K-means	0.2	0.75 (0.25)	0.25	0.48 (0.13)	0.99	0.37 (0.15)
K-median	**0.97**	**0.46 (0.2)**	**0.99**	**0.46 (0.19)**	**1.0**	**0.44 (0.2)**
K-MOM	**0.95**	**0.36 (0.11)**	**0.98**	**0.36 (0.12)**	**0.99**	**0.36 (0.12)**

Bibliographie

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study, *Annales de l'IHP Probabilités et statistiques*}, 48(4), pp. 1148-1185

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey, *Foundations of Computational Mathematics*, Springer, 19(5), pp. 1145-1190

Hampel, F. R. et Ronchetti, E. M. et Rousseeuw, P. J. and Stahel, W. A. (2011). Robust statistics: the approach based on influence functions, John Wiley & Sons, 196, pp. 98

Lecué, G. et Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice, *arXiv preprint arXiv:1711.10306*

Devroye, L. et Lerasle, M. et Lugosi, G. et Oliveira, R. I. et al. (2016). Sub-Gaussian mean estimators, *The Annals of Statistics*, Institute of Mathematical Statistics, 44(6), pp. 2695-2725

Al Hasan, M. and Chaoji, V. and Salem, S. and Zaki, M. J. (2009). Robust partitional clustering by outlier and density insensitive seeding, *Pattern Recognition Letters*, Elsevier, 30(11), pp. 994-1002

References

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11):994–1002, 2009.

[2] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[3] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[4] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[5] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Régression linéaire généralisée sur composantes supervisées pour les modèles à facteurs latents

Julien GIBAUD[1], Xavier BRY[1] et Catherine TROTTIER[1,2]

[1] Institut Montpelliérain Alexander Grothendieck, CNRS, Univ. Montpellier, France.
[2] Univ. Paul-Valéry Montpellier 3, F34000, Montpellier, France.

Contact : julien.gibaud@umontpellier.fr, xavier.bry@umontpellier.fr et catherine.trottier@univ-montp3.fr.

Résumé

À l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) a été conçue pour trouver des composantes explicatives conjointement supervisées par un ensemble de réponses au sein de très nombreuses covariables redondantes, ce qui est nécessaire dans un contexte de grande dimension. Dans ce travail, nous proposons d'étendre la méthode SCGLR dans l'objectif de modéliser la matrice de variance-covariance des réponses de telle sorte que la corrélation entre ces réponses soit principalement expliquée par quelques facteurs. Nous cherchons à identifier des blocs dans la matrice de variance-covariance pour les réponses partageant des dépendances mutuelles. Dans un cadre écologique par exemple, nous nous intéressons aux relations statistiques entre les présences de différentes espèces. Un algorithme basé sur EM est proposé afin d'estimer le modèle.

Mots clefs: SCGLR, modèle à facteurs, algorithme EM, variables latentes.

Abstract

Originally, the Supervised Component-based Generalized Linear Regression (SCGLR) was designed to find explanatory components jointly supervised by a set of responses in many redundant covariates, something much needed in a high-dimensional framework. In this work, we propose to extend the SCGLR methodology with the objective of modeling the responses variance-covariance matrix in such a way that the correlation between these responses is mainly explained by few factors. We aim at identifying blocks in the variance-covariance matrix depicting the outcomes sharing mutual dependencies. In an ecological framework for instance, we study the statistical relations between the presences of different species. An algorithm based on EM is proposed in order to estimate the model.

Keywords: SCGLR, factor model, EM algorithm, latent variables.

1 Contexte

Les changements climatiques entraînent certains dérèglements des écosystèmes pouvant causer des extinctions d'espèces animales ou végétales. Dans ce contexte, le développement de modèles permettant de prédire le futur de la biodiversité est devenu un enjeu crucial. Récemment, de nombreuses avancées ont été faites dans ce domaine, en particulier par l'extension des modèles de distribution des espèces (Species Distribution Models, SDM), qui traitent les espèces séparément, à des modèles de distribution jointe (Joint Species Distribution Models, JSDM). Les JSDM permettent de formaliser l'interdépendance des espèces et de mieux comprendre son impact sur la composition des communautés. Par ailleurs, modéliser l'abondance des espèces requiert de

prendre en compte un grand nombre de covariables explicatives souvent corrélées, ce qui impose une réduction de dimension et la régularisation des modèles.

Dans leur article, Bry et *al.* [1] proposent une méthode - la régression linéaire généralisée sur composantes supervisées (Supervised Component-based Generalized Linear Regression, SCGLR) - combinant le modèle linéaire généralisé multivarié avec les méthodes à composantes permettant la réduction de dimension. SCGLR optimise un critère compromis entre la qualité d'ajustement (Goodness-of-Fit, GoF) et la pertinence structurelle (Structural Relevance, SR) [2] mesurant la proximité des composantes supervisées à des dimensions d'intérêt. Cette technique ne trouve pas seulement des directions fortes et interprétables, elle produit aussi des prédicteurs régularisés, ce qui permet le traitement de données de grande dimension. Cependant, SCGLR suppose que les réponses sont indépendantes les unes des autres. Pour nous affranchir de cette hypothèse, nous proposons d'étendre cette méthode aux modèles à facteurs. L'objectif est de modéliser la matrice de variance covariance des réponses (espèces) afin d'identifier celles qui auraient des dépendances mutuelles.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, puis son extension aux modèles à facteurs latents.

2.1 SCGLR

N individus sont décrits par K réponses y^k, $k = 1, \ldots, K$, ainsi que des covariables explicatives séparées en deux groupes : un groupe X de covariables *a priori* nombreuses et possiblement redondantes, et un autre A de covariables additionnelles peu nombreuses et faiblement, voire non-redondantes. On notera $X \in \mathbb{R}^{N\times P}$ et $A \in \mathbb{R}^{N\times R}$ les matrices correspondantes. Chaque réponse y^k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM) [8]. Pour la partie explicative du modèle, seule la matrice X requiert réduction de dimension et régularisation. À cette fin, SCGLR cherche dans X des composantes communes à l'ensemble des réponses. Une composante $f \in \mathbb{R}^N$ est donnée par $f = Xu$ où $u \in \mathbb{R}^P$ est un vecteur de coefficients. Le prédicteur linéaire associé à la réponse y^k est donné par :

$$\eta^k = (Xu)\,\gamma^k + A\delta^k,$$

où γ^k et δ^k sont les paramètres de régression. La composante f est commune à l'ensemble des réponses y^k et pour assurer son identifiabilité, nous imposons $u^T M^{-1} u = 1$, où $M \in \mathbb{R}^{P\times P}$ est une matrice symétrique définie positive. Nous supposons que les réponses sont indépendantes conditionnellement aux variables explicatives.

À cause du produit $u\gamma^k$, le modèle "linéarisé" à chaque étape de l'algorithme des scores de Fisher (Fisher Scoring Algorithm, FSA) pour l'estimation du GLM, n'est pas linéaire et doit être estimé de façon alternée sur u et sur $\{\gamma^k, \delta^k\}$. Soient w^k, la pseudo-réponse (ou variable de travail) associée à chaque étape du FSA, et W_k^{-1} sa matrice de variance-covariance. L'estimateur des moindres carrés de u est solution des programmes équivalents suivants :

$$\min_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| w^k - \Pi^{W_k}_{\mathrm{vect}(f,A)} w^k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| \Pi^{W_k}_{\mathrm{vect}(f,A)} w^k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \psi_A(u),$$

avec $\psi_A(u) = \sum_{k=1}^{K} \left\|w^k\right\|_{W_k}^2 \cos^2_{W_k}\left(w^k, \Pi^{W_k}_{\text{vect}(f,A)} w^k\right)$. La quantité ψ_A est une mesure de GoF. Pour trouver des composantes fortes et interprétables, le GoF ne suffit pas. Il faut le combiner avec une mesure de pertinence structurelle (SR).

Dans ce travail nous utilisons une mesure particulière de SR : l'inertie duale généralisée (Variable Powered Inertia, VPI). On appelle W la matrice des poids *a priori* des observations (typiquement, $W = \frac{1}{N} I_N$) et on suppose les colonnes de X centrées et réduites. Nous voulons trouver une direction vect(u) proche d'un faisceau. En un mot, un faisceau est un ensemble de variables explicatives suffisamment corrélées pour être vues comme alignées autour de la même dimension latente. Pour cela, on pose $l \geq 1$ et la SR s'écrit :

$$\phi(u) = \left(\frac{1}{P}\sum_{p=1}^{P}\left(u^T X^T W x^p x^{pT} W X u\right)^l\right)^{1/l} = \left(\frac{1}{P}\sum_{p=1}^{P}\langle Xu, x^p\rangle_W^{2l}\right)^{1/l}.$$

Le paramètre l permet de trouver une composante proche d'un faisceau plus (l fort) ou moins (l faible) étroit de variables corrélées. De façon générale, quelle que soit la SR choisie, la métrique M de la contrainte $u^T M^{-1} u = 1$ est écrite de la forme $M^{-1} = \tau I_N + (1-\tau) X^T W X$, où $\tau \in [0,1]$ est un paramètre de régularisation de type ridge [6].

Pour construire un compromis entre le GoF et la SR, SCGLR introduit un réel $s \in [0,1]$ traduisant leur poids respectif et considère le programme de maximisation suivant :

$$\max_{u, u^T M^{-1} u = 1} \phi(u)^s \psi_A(u)^{1-s} \Leftrightarrow \max_{u, u^T M^{-1} u = 1} s \ln(\phi(u)) + (1-s)\ln(\psi_A(u)). \tag{1}$$

Afin de trouver les composantes d'ordre $h > 1$, nous notons $f^h = Xu^h$ la h-ième composante et $F^h = [f^1, \ldots, f^h]$ la matrice des h premières composantes, avec $h < H$. Par simplification, nous notons F la matrice des H composantes. Nous adoptons alors le principe d'emboîtement local (Local Nesting, LocNes) présenté par [3]. Suivant ce principe, la composante supplémentaire f^{h+1} doit venir compléter au mieux les composantes précédentes en plus de la matrice A, c'est à dire $A^h := [F^h, A]$. Ainsi, f^{h+1} est calculée en utilisant A^h comme matrice de covariables additionnelles. De plus, nous imposons que f^{h+1} soit orthogonale à F^h par la contrainte ${F^h}^T W f^{h+1} = 0$. Cette maximisation sous contrainte est permise par l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING) [7].

2.2 SCGLR pour modèle à facteurs

À l'origine, la méthode SCGLR fut développée dans un contexte de modèle linéaire généralisé, cependant, dans ce travail, nous nous limiterons à une matrice $Y \in \mathbb{R}^{N \times K}$ de réponses gaussiennes. Nous appelons y_n, f_n et a_n les vecteurs composés de la n-ième ligne des matrices Y, F et A respectivement. Chaque y_n est expliquée par f_n, a_n et par L variables latentes $g_n = (g_n^1, \ldots, g_n^L)^T$ appelées facteurs. Ainsi, le modèle s'écrit :

$$\underbrace{y_n}_{K\times 1} = \underbrace{\Gamma^T}_{K\times H}\underbrace{f_n}_{H\times 1} + \underbrace{\Delta^T}_{K\times R}\underbrace{a_n}_{R\times 1} + \underbrace{B^T}_{K\times L}\underbrace{g_n}_{L\times 1} + \underbrace{\varepsilon_n}_{K\times 1}, \tag{2}$$

où $\Gamma = [\gamma^1, \ldots, \gamma^K]$, $\Delta = [\delta^1, \ldots, \delta^K]$ et $B = [b^1, \ldots, b^K]$ sont des paramètres de régression et où $\varepsilon_n \sim \mathcal{N}(0, \Psi)$ avec $\Psi = \text{diag}(\sigma_k^2)_{k=1,\ldots,K}$, représente les erreurs indépendantes. De plus g_n est supposé suivre une loi $\mathcal{N}(0, I_L)$ et être indépendant de ε_m pour toutes valeurs de n et m. Ces hypothèses impliquent que le modèle est construit de telle manière que toute la corrélation entre

les réponses soit expliquée par les L facteurs.

Afin d'estimer les paramètres, nous devons maximiser la log-vraisemblance du modèle $l(\Theta; Y)$, où $\Theta = \{\Gamma, \Delta, B, \Psi\}$. Cependant, à cause des facteurs non observés, cette log-vraisemblance possède une expression complexe qui la rend difficile à maximiser. Ainsi, nous utilisons l'algorithme EM [4] pour estimer les paramètres. L'étape M de l'algorithme consiste à maximiser l'espérance conditionnelle de la log-vraisemblance complétée $\mathbb{E}[l(\Theta; Y, G)|Y; \Theta']$, cette espérance étant mise à jour dans l'étape E.

2.2.1 Étape E (Espérance conditionnelle)

Pour réaliser l'étape E de l'algorithme, nous devons calculer explicitement l'espérance conditionnelle de la log-vraisemblance complétée. Cette dernière s'écrit :

$$\mathbb{E}[l(\Theta; Y, G)|Y, \Theta'] = \sum_{n=1}^{N} \int \ln\left(f(y_n|g_n; \Theta) f(g_n; \Theta)\right) f(g_n|y_n; \Theta') dg_n.$$

Ainsi, nous devons préalablement calculer les lois de $y_n|g_n$ et de $g_n|y_n$. La loi de $y_n|g_n$ est donnée par le modèle (2) tandis que la loi de $g_n|y_n$ est donnée par la règle de l'espérance conditionnelle des lois Gaussiennes multivariées :

$$\text{Si} \quad \begin{pmatrix} y_n \\ g_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \Gamma^T f_n + \Delta^T a_n \\ 0 \end{pmatrix}, \begin{pmatrix} B^T B + \Psi & B^T \\ B & I_L \end{pmatrix}\right),$$

alors $g_n|y_n \sim \mathcal{N}(\alpha(y_n - \Gamma^T f_n - \Delta^T a_n), I_L - \alpha B^T)$, où $\alpha = B(B^T B + \Psi)^{-1}$. Finalement, tout calcul fait, l'espérance de la log-vraisemblance complétée devient :

$$\begin{aligned}\mathbb{E}[l(\Theta; Y, G)|Y, \Theta'] = -\frac{1}{2}\Bigg\{ & N(K+L)\ln(2\pi) + N\sum_{k=1}^{K}\ln(\sigma_k^2) + \sum_{n=1}^{N}\mathbb{E}\left[g_n^T g_n|y_n; \Theta'\right] + \\ & \sum_{k=1}^{K}\frac{1}{\sigma_k^2}\left[\left\|y^k - F\gamma^k - A\delta^k\right\|^2 + b^{k\,T}\tilde{R}b^k - 2\left(\tilde{G}b^k\right)^T\left(y^k - F\gamma^k - A\delta^k\right)\right]\Bigg\},\end{aligned}$$

où les lignes de la matrice $\tilde{G}$ sont les moments d'ordre 1 :

$$\tilde{g}_n := \mathbb{E}(g_n|y_n; \Theta) = \alpha(y_n - \Gamma^T f_n - \Delta^T a_n) \tag{3}$$

et où $\tilde{R} = \sum_{n=1}^{N}\tilde{R}_n$ est la somme des moments d'ordre 2 :

$$\tilde{R}_n := \mathbb{E}(g_n g_n^T|y_n; \Theta) = \mathbb{V}(g_n|y_n; \Theta) + \mathbb{E}(g_n|y_n; \Theta)\mathbb{E}(g_n|y_n; \Theta)^T = I_L - \alpha B^T + \tilde{g}_n\tilde{g}_n^T. \tag{4}$$

2.2.2 Étape M (Maximisation)

Dans un objectif d'identification, nous avons besoin de contraindre la matrice B. Comme démontré par [5], si Ω est une matrice orthogonale, nous pouvons remplacer le produit $B^T g_n$ par $B_0^T g_{0n}$ dans lequel le nouveau facteur $g_{0n} = \Omega g_n$ est une rotation de l'ancien facteur g_n. Les conditions sur les moments respectées par les anciens facteurs sont aussi respectées par les nouveaux, autrement dit, $\mathbb{E}(g_{0n}) = \Omega\mathbb{E}(g_n) = 0$ et $\mathbb{V}(g_{0n}) = \Omega\mathbb{V}(g_n)\Omega^T = I_L$. De plus, les paramètres aussi subissent une rotation. Les nouveaux paramètres sont liés aux anciens par $B_0^T = B^T\Omega^T$. Étant donné que ces nouveaux paramètres et facteurs donnent lieu à la même distribution, ils ne peuvent être identifiés

à partir des observations que si des restrictions supplémentaires sont imposées. Nous choisissons d'imposer la matrice B contrainte :

$$B = \begin{pmatrix} b_1^1 & \dots & b_1^L & b_1^{L+1} & \dots & b_1^K \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & b_L^L & b_L^{L+1} & \dots & b_L^K \end{pmatrix},$$

où pour tout $k < l$, $k, l = 1, \dots, L$, $b_l^k = 0$ et où pour tout $l = 1, \dots, L$, $b_l^l > 0$.

La contrainte n'imposant rien sur les paramètres Γ, Δ et Ψ, nous maximisons normalement sur ces derniers. Ainsi, pour tout $k = 1, \dots, K$, la maximisation de $\mathbb{E}[l(\Theta; Y, G)|Y, \Theta']$ sur $\{\gamma^k, \delta^k, \sigma^k\}$ donne :

$$\begin{pmatrix} \hat{\gamma}^k \\ \hat{\delta}^k \end{pmatrix} = \left([F, A]^T [F, A]\right)^{-1} [F, A]^T \left(y^k - \tilde{G} b^k\right), \tag{5}$$

$$\hat{\sigma}_k^2 = \frac{1}{N} \left\{ \left\| y^k - F\gamma^k - A\delta^k \right\|^2 + b^{k\,T} \tilde{R} b^k - 2 \left(\tilde{G} b^k\right)^T \left(y^k - F\gamma^k - A\delta^k\right) \right\}. \tag{6}$$

Désormais, nous devons maximiser sur le vecteur b^k sous la contrainte. Pour tout $k = 1, \dots, L$, posons $b^k = (b_{1:k}^{k\,T}, \mathbf{0}^T)^T$, où $b_{1:k}^k = (b_1^k, \dots, b_k^k)^T$ est un vecteur de longueur k et $\mathbf{0}$ le vecteur nul de longueur $L - k$. Ainsi, après maximisation, on a pour tout $k = 1, \dots, L$,

$$\hat{b}_{1:k}^k = \left(\tilde{R}_{1:k}^{1:k}\right)^{-1} \left(\tilde{G}^{1:k}\right)^T \left(y^k - F\gamma^k - A\delta^k\right), \tag{7}$$

où $\tilde{R}_{1:k}^{1:k}$ est la sous matrice de taille $k \times k$ de $\tilde{R}$ et où $\tilde{G}^{1:k}$ est la matrice composée des k premières colonnes de $\tilde{G}$. De la même manière, pour $k = L + 1, \dots, K$, on a :

$$\hat{b}^k = \tilde{R}^{-1} \tilde{G}^T \left(y^k - F\gamma^k - A\delta^k\right). \tag{8}$$

2.2.3 Contrainte sur le nombre maximal de facteurs

Nous appelons $\Sigma = B^T B + \Psi$ la matrice de variance-covariance de y_n. Il existe seulement $K(K+1)/2$ éléments distincts dans Σ, cependant il y a LK éléments dans B plus K éléments dans Ψ. La contrainte impose *a priori* $L(L-1)/2$ éléments nuls sur la matrice B. Finalement, $B^T B + \Psi$ possède $LK + K - L(L-1)/2$ éléments distincts. Pour déterminer ces paramètres, nous devons avoir $K(K+1)/2 \geq LK + K - L(L-1)/2$ ou, autrement dit, $L \leq (2K + 1 - \sqrt{8K+1})/2$. Nous donnons quelques exemples de nombres maximaux de facteurs en fonction du nombre de réponses :

K	1	2	3	4	5	6	7	8	9	10
L max	0	0	1	1	2	3	3	4	5	6

Table 1: Nombre maximal de facteurs

En pratique, pour simplifier la structure de variance-covariance résiduelle des réponses, le nombre de facteurs restera faible.

3 Algorithme

Des essais numériques, sur données simulées et réelles, impliquant l'Algorithme 1 seront exposés lors de la présentation orale de ces travaux de recherche.

Algorithm 1: SCGLR pour les modèles à facteurs

while *not convergence* **do**

Répéter les étapes (3) et (4) puis les étapes (5), (6), (7) et (8) jusqu'à la convergence de l'algorithme EM

$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\ l(\Theta^{(t)}; Y)$

Répéter sur le nombre de composantes la maximisation du critère (1) à l'aide de l'algorithme PING

$\forall h = 1, \ldots, H, \quad f^{h\,(t+1)} = X u^{h\,(t+1)}$

$t \leftarrow t + 1$

end

Remerciements

Cette recherche a été soutenue par le projet GAMBAS financé par l'Agence Nationale de la Recherche (ANR-18-CE02-0025).

References

[1] Xavier Bry, Catherine Trottier, Thomas Verron, and Frédéric Mortier. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119: 47–60, 2013.

[2] Xavier Bry and Thomas Verron. THEME: THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12): 637–647, 2015.

[3] Xavier Bry, Thomas Verron, and Pierre Cazes. Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression. *Analytica chimica acta*, 642(1-2): 45–58, 2009.

[4] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

[5] John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2): 557–587, 1996.

[6] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1): 69–82, 1970.

[7] Alston S. Householder. *The theory of matrices in numerical analysis*. Courier Corporation, 2013.

[8] P. McCullagh and J.A. Nelder. 1989, Generalized Linear Models, Chapman and Hall, New York, NY.

Programme de la Session Jeunes

1. Présentation des associations utiles pour les doctorants
   1. En France : Groupe Jeunes Statisticien.ne.s de la SFdS, présentation des activités par Geneviève Robin
   2. En Europe : “Young Statisticians Europe” et YoungStats, présentation par Andrej Srakar
   3. Postuler avec sérénité : les tips and tricks d’opération postes, présentation par Geneviève Robin

2. Table ronde : Quelles opportunités après la thèse ?
Cinq docteurs témoignent de leurs parcours, en France ou à l’étranger, dans la recherche académique ou industrielle.
* Christian Capezza, Post-doc, Université de Naples, Italie
* Benjamin Guedj, Principal Researcher (University College London) and Researcher (Inria)
* Christine Keribin, Maître de Conférences Hors-Classe, Université Paris-Saclay
* Claire Roman, Consultante scientifique et Data scientist, Rayce-Araymond
* Alkéos Michail, Chief Technology Officer, AgenT-biotech, Paris

Décomposition d'une somme aléatoire via une approche Bayésienne approximative

Pierre-O Goffard [1]

[1] *Institut de Science Financière et Assurances*
50 Avenue Tony Garnier, 69007 Lyon
pierre-olivier.goffard@univ-lyon1.fr

Résumé. Soit la variable aléatoire (v.a.)

$$X = \sum_{i=1}^{N} U_i,$$

où N est une v.a. de comptage et les $(U_i)_{i\geq 1}$ forment une suite de v.a. continues et positives. En assurance, X représente le montant total des sinistres sur une période d'exercice donnée. La v.a. N correspond au nombre de sinistres et les $(U_i)_{i\geq 1}$ sont les indemnisations associées à chaque sinistre. Un modèle paramétrique est considéré pour la fréquence et le montant des sinistres, l'objectif de ce travail est de proposer une méthode d'inférence des paramètres sur la base d'un échantillon IID distribué comme X. Une approche Bayésienne approximative est adoptée pour s'affranchir de l'absence d'une formule fermée pour la fonction de vraisemblance du modèle.

Mots-clés. Distributions composées, Actuariat, Approximate Bayesian Computation.

Abstract. Consider the random variable (r.v.)

$$X = \sum_{i=1}^{N} U_i,$$

where N is a conting r.v. and the U_i's form a sequence of nonnegative r.v.. In insurance, X represent the total claim sizes over a given time period. The r.v. N corresponds to the number of claims and the sequence $(U_i)_{i\geq 1}$ are the compensations associated to each claim. The claim frequency and the claim amounts are governed by a parametric model and the goal is draw inference based on an IID sample $x_1, \ldots, x_t$ distributed like X. The absence of a tractable likelihood function is overcome by using a method called Approximate Bayesian Computation.

Keywords. Compound distribution, Actuarial Science, Approximate Bayesian Computation.

Description de la communication

Sur une période d'exercice donnée, un nombre aléatoire de sinistres sont reportés à une compagnie d'assurance, et chaque sinistre est associé à une indemnisation dont le montant est aléatoire. La fréquence des sinistres est une variable aléatoire de comptage alors que les montants indemnisés sont des variables aléatoires positives et continues. La distribution de la fréquence et des montants est caractérisée par des paramètres $\boldsymbol{\theta}_{\text{freq}}$ et $\boldsymbol{\theta}_{\text{sev}}$ respectivement, avec $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{freq}}; \boldsymbol{\theta}_{\text{sev}})$. Pour chaque période $s = 1, \ldots, t$, le nombre de sinistres n_s et les montants indemnisés $\boldsymbol{u}_s \coloneqq (u_{s,1}, u_{s,2}, \ldots, u_{s,n_s})$ sont modélisés via

$$n_s \sim p_N(n\,;\, \boldsymbol{\theta}_{\text{freq}}), \quad \text{et} \quad (\boldsymbol{u}_s \mid n_s) \sim f_U(\boldsymbol{u}\,;\, n, \boldsymbol{\theta}_{\text{sev}}).$$

La calibration de ces modèles est centrale dans la gestion des risques de la compagnie d'assurance. Le coût moyen des sinistres est par exemple utilisé pour fixer le niveau des primes versées par les assurés. La nature mixte des données de sinistralité, comprenant une composante discrète et une composante continue, a conduit à l'élaboration de deux approches. La première prend en compte séparément la fréquence et le coût, voir [7]. La deuxième rassemble les deux composantes au sein d'un modèle composée pour lequel les données sous format agrégé suffisent à la calibration. La deuxième approche est adoptées dans le cadre de cette communication car les données individuelles $\{(n_1, \boldsymbol{u}_1), \ldots, (n_t, \boldsymbol{u}_t)\}$ ne sont pas accessibles. Seuls les montant agrégés,

$$x_s = \sum_{i=1}^{n_s} u_{s,i} \tag{1}$$

pour chaque période d'exercice $s = 1, \ldots, t$ sont disponibles. La pratique actuarielle suppose en générale que la fréquence suit une loi de Poisson et que les montants indemnisés suivent une loi gamma [10]. Ce modèle, dit de Tweedie [13], est couramment utilisé pour la tarification [12], et pour le provisionnement [15]. Le modèle de Tweedie est aussi populaire pour l'étude des précipitations [6]. L'objectif est d'aller au delà du modèle de Tweedie, pour pouvoir considérer des distributions autre que Poisson et gamma mais aussi pour inclure de la dépendance soit entre les montants, soit entre la fréquence et les montants. L'objectif est de trouver la valeur des paramètres $\boldsymbol{\theta}_{\text{freq}}$ et $\boldsymbol{\theta}_{\text{sev}}$ qui permet le meilleur ajustement du modèle paramétrique considéré aux données $\boldsymbol{x} = (x_1, \ldots, x_t)$.

Les données considérés (1) peuvent être vus comme les incréments d'un processus stochastique $(Z_t)_{t\geq 0}$ défini par

$$Z_t = \sum_{i=1}^{N_t} U_i, \ t \geq 0, \tag{2}$$

observé à intervalle de temps régulier. Le processus $(N_t)_{t\geq 0}$ est un processus de comptage et la suite $(U_i)_{i\geq 1}$ est une suite de v.a. positives. En théorie du risque, le processus $(Z_t)_{t\geq 0}$

représente les engagements de la compagnie d'assurance envers les assurés à une date $t \geq 0$ [1]. Une méthode de décomposition de somme aléatoire [3] construit un estimateur non paramétrique de la distribution des sauts en se basant uniquement sur des observations de la trajectoire de $(Z_t)_{t\geq 0}$. Ce problème a été considéré par différents auteurs [14, 4, 9] avec des applications notamment en théorie des files d'attente. La méthodologie qui fait l'objet de cette communication propose de décomposer une somme aléatoire dans un cadre paramétrique.

En statistique Bayésienne, le paramètre $\boldsymbol{\theta}$ est une variabale aléatoire et son estimation repose sur l'étude de sa loi *a posteriori* $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Le théorème de Bayes permet d'accéder à la loi *a posteriori*

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{3}$$

via la mise à jour de la loi *a priori* $\pi(\boldsymbol{\theta})$ par la fonction de vraiemblance $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ [8]. La loi *a posteriori* (3) admet rarement une forme explicite et est approchée par une distribution empirique. Des échantillons de loi *a posteriori* sont obtenus par le biais d'algorithme de simulation de Monte Carlo par chaine de Markov (MCMC). L'échantillonage par des algorithmes MCMC nécessite d'être en mesure d'évaluer la vraisemblance des données $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Au vu de la définition de X, donnée dans l'équation (1), il y a très peu de cas qui conduisent à une formule simple pour la vraisemblance.

Le calcul Bayésien approximatif (ABC) [11] permet de s'affranchir de la vraisemblance à la condition d'être en mesure de générer des données artificielles à partir du modèle considéré. Cette technique a attiré beaucoup d'attention récemment en capitalisant sur un large champ d'application et un mode de fonctionnement très intuitifs. L'algorithme répète les étapes suivantes

(i) Tirage aléatoire du paramètre $\boldsymbol{\theta}$ depuis la loi *a priori* $\tilde{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta})$;

(ii) simulation de données artificielles $\tilde{\boldsymbol{x}}$ suivant le modèle considéré $(\tilde{\boldsymbol{x}} \mid \tilde{\boldsymbol{\theta}}) \sim p(\boldsymbol{x} \mid \boldsymbol{\theta})$;

(iii) Si $\mathcal{D}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \epsilon$, alors la valeur du paramètre $\tilde{\boldsymbol{\theta}}$ est conservée,

où $\mathcal{D}(\,\cdot\,,\,\cdot\,)$ est une mesure de dissimilarité et $\epsilon > 0$ désigne un seuil de tolérance. L'algorithme retourne un échantillon $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$, dont la distribution empirique approche la loi *a posteriori* $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$.

L'algorithme ABC présenté dans cet exposé repose sur la distance de Wasserstein [2] et un échantilloneur de Monte Carlo séquentiel [5]. Il permet de considérer toute sorte de distribution pour les montants et la fréquence des sinistres ainsi que de s'affranchir de certaines hypothèses classiques comme le fait que les montants soientt IID et indépendants de la fréquence.

References

[1] Søren Asmussen and Hansjörg Albrecher. *Ruin Probabilities*, volume 14 of *Advanced Series on Statistical Science and Applied Probability*. World Scientific, 2nd edition, 2010.

[2] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 10 2019.

[3] Boris Buchmann and Rudolf Grübel. Decompounding: an estimation problem for Poisson random sums. *Ann. Statist.*, 31(4):1054–1074, 08 2003.

[4] Alberto J Coca. Efficient nonparametric inference for discretely observed compound Poisson processes. *Probability Theory and Related Fields*, 170(1-2):475–523, 2018.

[5] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[6] Peter K. Dunn. Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24(10):1231–1239, jul 2004.

[7] Edward W. Frees. Frequency and severity models. In Edward W. Frees, Richard A. Derrig, and Glenn Meyers, editors, *Predictive Modeling Applications In Actuarial Science*, pages 138–164. Cambridge University Press.

[8] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[9] Shota Gugushvili, Frank van der Meulen, and Peter Spreij. A non-parametric Bayesian approach to decompounding from high frequency data. *Statistical Inference for Stochastic Processes*, 21(1):53–79, 2018.

[10] Bent Jørgensen and Marta C. Paes De Souza. Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93, jan 1994.

[11] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

[12] Gordon K. Smyth and Bent Jørgensen. Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1):143–157, 2002.

[13] Maurice CK Tweedie. An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604, 1984.

[14] Bert van Es, Shota Gugushvili, and Peter Spreij. A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13(3):672–694, 08 2007.

[15] Mario V. Wüthrich. Claims reserving using tweedie's compound poisson model. *ASTIN Bulletin*, 33(2):331–346, nov 2003.

Multiple Co-clustering de séries temporelles. Application à la validation de systèmes d'aide à la conduite

Etienne Goffinet [1,2] & Mustapha Lebbah [1] & Hanane Azzag [1] & Loïc Giraldi [2]

[1] *Université Sorbonne Paris Nord, CNRS, Laboratoire Informatique Paris Nord, 93430 Villetaneuse*
[2] *Groupe Renault, 78280 Guyancourt*

Résumé. Le développement de systèmes d'aide à la conduite demeure un défi technique pour les constructeurs automobiles. La validation de ces systèmes nécessite de les éprouver dans un nombre considérable de contextes de conduites. Pour ce faire, le Groupe Renault a recourt à la simulation massive, qui permet de reproduire précisément la complexité des conditions physiques de conduite et produit une grande quantité de séries temporelles multivariées. Nous présentons les contraintes opérationnelles et défis scientifiques liées à ces jeux de données, ainsi que notre proposition d'une approche de classification probabiliste adaptée, qui crée plusieurs partitions indépendantes en regroupant les variables redondantes. Nous présentons les résultats obtenus avec cette nouvelle méthode sur un jeu de données issu d'un cas d'usage industriel.

Mots-clés. Multi-Coclustering, Séries temporelles multivariées, Multi-Clustering

Abstract.
Advanced driver-assistance systems development remains a technical challenge for automobile manufacturers. The reliable validation of these systems requires testing them in a considerable number of driving contexts. The numerical simulation helps Groupe Renault validate such systems, and recreates the complexity of physical driving conditions. These simulations produce large quantities of high-dimensional multivariate time series. We detail the operational constraints associated to these datasets, and a suited model-based classification method able to handle. Based on a hierarchical structure, it creates several independent partitions while grouping redundant variables. We present the results obtained on a dataset from an industrial use case.

Keywords. Multi-Coclustering, Multivariate time series, Multi-Clustering

1 Introduction

Avant de pouvoir être mis sur le marché, les systèmes d'aide à la conduite sont rigoureusement étudiés et testés par les constructeurs automobiles. Pour garantir la qualité de ces tests, fonction de leur exhaustivité, le Groupe Renault a recourt à la simulation massive. Pour un cas d'usage donné, cette validation produit un nombre important

d'observations (en centaines de milliers) décrites par un grand nombre de capteurs (en centaines). L'exploration de ces données permet de déterminer précisément les capacités d'un système d'aide à la conduite et de raffiner son paramétrage par l'expert. Le comportement de la voiture simulée ne peut pas toujours être prédit, ce qui rend le recours aux méthodes non-supervisées indispensable. La classification non-supervisée, ou *clustering*, regroupe des observations similaires en *clusters*. Le clustering de séries temporelles basé sur des modèles (Goffinet (2020)) nous est particulièrement utile, en particulier la production d'intervalles de confiance pour la détection probabiliste des valeurs aberrantes et la production d'intervalles de confiances sur les paramètres inférés. Certaines des variables obtenues en sortie de simulation sont corrélées (par exemple la vitesse d'une voiture et la vitesse de rotation d'une de ses roues), positivement ou négativement, et parfois même dupliquées. D'autres, bien que non-dupliquées et non-corrélées, produisent des partitions similaires lorsqu'elles sont traitées individuellement par une méthode de partitionnement donnée (par exemple position et accélération). Nous proposons une nouvelle méthode qui regroupe les variables de manière hiérarchique: basé sur leur partition ligne, puis sur leur distribution. Le co-clustering (Figure 1 - a)) réalise simultanément un clustering d'observations (aussi appelé partition ligne dans la suite) et un clustering de variables (aussi appelé partition colonne dans la suite). Ce modèle fait l'hypothèse que toutes les variables partagent la même partition ligne. Lorsque les variables ne satisfont pas cette hypothèse de partage de partition ligne (e.g. les variables décrivent des physiques ou des systèmes différents), le co-clustering n'est plus adapté.

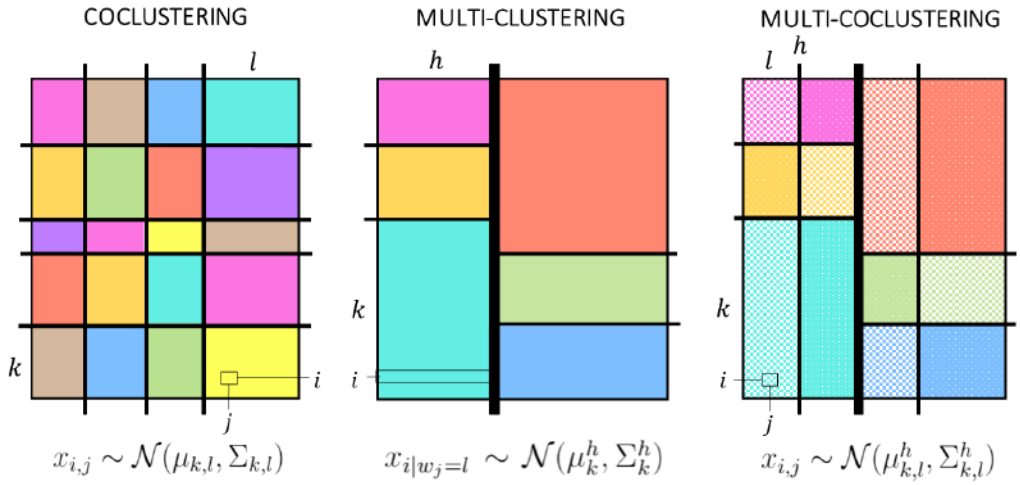


Figure 1: Illustration du Co-clustering, du Multi-Clustering et du Multi-Coclustering. k, l, h sont les indices d'un cluster d'observations, de variables corrélées et de variables redondantes (respectivement). Couleurs et motifs indiquent les appartenances aux blocs.

Cette contrainte peut être relâchée avec le *Multi-Clustering* ((Hu, J. (2018), Marbac

(2019), Vandewalle, V. (2020)) où plusieurs partitions lignes différentes sont inférées, comme illustré sur la Figure 1 - b). Dans ce cas, les partitions ligne ne sont plus mélangées, mais définies indépendamment dans chaque cluster de variables. Dans un contexte paramétrique, la combinatoire liée au Multi-Clustering rend difficile la sélection de modèle par recherche exhaustive (il y a, par exemple, 184755 modèles différents avec au maximum 10 clusters de variables et 10 clusters d'observations). Des approches heuristiques sont alors envisagées, (par exemple des stratégies gloutonnes) au prix d'hypothèses sur la structure du modèle. L'approche non-paramétrique permet de contourner cette contrainte en intégrant nativement une sélection de modèle. Un modèle de Multi-Clustering Bayésien Non-paramétrique a été développée par Guan (2010), pour le traitement de données continues multivariées. Dans ce modèle, les lignes appartenant à un bloc suivent une distribution multivariée indépendante (c.f. Figure 1-b)). Cette méthode de Multi-Clustering regroupe les variables en fonction de leur partition ligne, mais ne regroupe pas les variables de distribution similaire. Nous proposons dans ce papier une nouvelle méthode de Multi Co-Clustering (MCC) non-paramétrique qui regroupe les variables partageant la même partition d'observations et, dans chaque groupe de variables, infère un modèle de blocs latents non-paramétriques (NPLBM) (Meeds (2010)). Cette approche permet de discriminer plus finement les variables : parmi celles qui partagent les mêmes partitions lignes, la méthode regroupe les variables ayant des distributions identiques (c.f. Figure 1-c)). À notre connaissance, il n'existe qu'un seul travail comparable sur le sujet par Tokuda (2017) mais qui ne s'applique pas aux séries temporelles.

2 Multi-Coclustering de séries temporelles multivariées

Dans cette partie nous définissons le modèle (Section 2.1), son inférence (Section 2.2) et présentons une application sur un jeu de donnée issu d'un cas d'usage industriel.

2.1 Définition

Chaque cellule du jeu de données final est un vecteur de coefficients issu d'une PCA fonctionnelle appliquée aux séries exprimées en base polynomiale, traitement classique dans le cas du co-clustering fonctionnel, (Bouveyron (2018), Slimen (2018)). Dans la suite, X désigne le dataset complet, de dimension $n \times p \times d$, avec n le nombre de simulations, p le nombre de variables, d la dimension de l'espace des observations. Soit H le nombre total de clusters de variables, tel que les variables dans un cluster partagent la même partition ligne. Ces clusters seront indicés par la variable $h \in \{1, \dots, H\}$. Pour chacun de ces clusters, soit p_h le nombre de clusters de variables possédant une distribution identique. Le vecteur $\mathbf{v} \in \mathbb{N}^p$ représente les appartenances aux clusters de variables partageant une même partition ligne, et le vecteur $\mathbf{w}_h \in \mathbb{N}^{p_h}$ l'appartenance aux clusters de variables

distribution identique. Ainsi, la paire $(\mathbf{v}, \mathbf{w}_h)$ est nécessaire pour identifier l'appartenance d'une variable à un cluster de colonnes. La matrice $Z \in \mathbb{N}^{H \times n}$ désigne les appartenances aux partitions lignes. Le modèle est défini par

$$x_{i,j} \mid v_j, w_j, z_{i,v_j}, \theta^{v_j}_{z_{i,v_j},w_j} \sim \mathcal{N}(\theta^{v_j}_{z_{i,v_j},w_j}),\ \theta^{v_j}_{z_{i,v_j},w_j} \sim G_0$$

$$v_j \sim Mult(\eta),\ \eta_j(\mathbf{r}) = r_j \prod_{j'=1}^{j-1} (1 - r_{j'}),\quad r_j \overset{\text{i.i.d.}}{\sim} Beta(1, \gamma)\ , \gamma \sim Gamma(a_r, b_r)$$

$$w_j \sim Mult(\rho_h),\ \rho^h_j(\mathbf{s}^h) = s^h_j \prod_{j'=1}^{j-1} \left(1 - s^h_{j'}\right),\quad s^h_j \overset{\text{i.i.d.}}{\sim} Beta(1, \beta_h)\ , \beta_h \sim Gamma(a_c, b_c)$$

$$z^{v_j}_j \sim Mult(\pi_h),\ \pi^h_j(\mathbf{t}^h) = t^h_j \prod_{j'=1}^{j-1} \left(1 - t^h_{j'}\right),\quad t^h_j \overset{\text{i.i.d.}}{\sim} Beta(1, \alpha_h),\ \alpha_h \sim Gamma(a_l, b_l).$$

Dans chaque dimension $(\mathbf{v}, W, Z)$, les proportions des appartenances suivent une construction *stick-breaking* (Sethuraman (1994)) dont le paramètre de concentration suit une loi Gamma. Les paramètres des distributions gaussiennes multivariées de chaque bloc (dont l'ensemble est noté Θ) suivent un prior conjugué Normal-inverse-Wishart, notée G_0. Dans la suite, on note $\chi = (G_0, a_r, b_r, a_c, b_c, a_l, b_l)$ l'ensemble des hyper-paramètres.

2.2 Inférence

Nous proposons une inférence en deux étapes: d'abord une étape de Multi-Clustering qui regroupe les variables partageant la même partition ligne, puis l'inférence d'un modèle NPLBM dans chacun de ces clusters.

Multi-Clustering Dans cette première étape, la distribution postérieure $p(\mathbf{v}, Z \mid X, \chi)$ est approchée par un échantillonneur de Gibbs. Chaque itération de l'algorithme se décompose en trois étapes: 1) Échantillonnage de $\mathbf{v}$ sachant Z; 2) Pour $h = \{1, \ldots, H\}$, mise à jour de Z^h sachant $\mathbf{v}$; 3) Mise à jour des paramètres de concentration. Dans l'étape 1), pour $j = 1, \ldots, p$, v_j est échantillonnée selon une loi multinomiale de proportions

$$p(v_j \mid \mathbf{v}_{-j}, Z, \chi) \propto \begin{cases} \dfrac{p_h}{p - 1 + \gamma}\, p(\mathbf{x}_{.,j} \mid \mathbf{z}^h, \chi), & \text{cluster existant } h, \quad (1) \\ \dfrac{\gamma}{n - 1 + \gamma}\, \mathrm{p}(\mathbf{x}_{.,j} \mid \chi), & \text{nouveau cluster}, \quad (2) \end{cases}$$

avec $\mathbf{v}_{-j} = \{v_i : i \neq j\}$ et $p(\mathbf{x}_{.,j} \mid \mathbf{z}^h)$ la distribution prédictive a priori de $x_{.,j}$ dans $\mathbf{z}^h$. Dans l'équation (2), $\mathrm{p}(\mathbf{x}_{.,j} \mid \chi)$ est estimé par inférence d'un modèle de mélange de processus de Dirichlet (DPM), (une fois par variable, avant l'inférence du MCC), qui produit également la partition ligne optimale $\hat{\mathbf{z}}^j$ associée au nouveau cluster échantillonné. Dans l'étape 2), les appartenances $\mathbf{z}^h$ sont mis à jour par un échantillonneur de Gibbs multivarié. L'étape 3) met à jour γ et α_h suivant la procédure définie par West (1998).

Coclustering Pour $h = \{1, \ldots, H\}$, un modèle NPLBM est inféré, à $\mathbf{z}^h$ fixé, par un algorithme de Gibbs approchant $p(\mathbf{w}^h \mid \mathbf{v}, \mathbf{z}^h)$. Pour chaque variable $j' = 1, \ldots, p_h$ du jeu de données $X^h = (x_{i,j} : v_j = h)$, une nouvelle appartenance est échantillonnée selon :

$$p(w^h_{j'} \mid \mathbf{w}^h_{-j'}, \mathbf{z}^h_{-j'}, \chi) \propto \begin{cases} \dfrac{p^h_l}{p_h - 1 + \beta} \displaystyle\prod_{k=1}^{K^h} p(\mathbf{x}^h_{k,j'} \mid \mathbf{x}^h_{k,l}, \chi), & \text{cluster existant } l, \quad (3) \\ \dfrac{\beta}{p_h - 1 + \beta} \displaystyle\prod_{k=1}^{K^h} p(\mathbf{x}^h_{k,j'} \mid \chi), & \text{nouveau cluster}, \quad (4) \end{cases}$$

avec $\mathbf{w}^h_{-j'}$ et $\mathbf{z}^h_{-j'}$ les appartenances aux blocs dans le cluster h sans la variable j', $q^h_l = \sum_{i \neq j'} \mathbb{I}_l(w_i)$, $\mathbf{x}^h_{k,j'} = \{x_{i,j'} : v_{j'} = h, z^h_i = k\}$, et $p(\mathbf{x}^h_{k,j'} \mid \mathbf{x}^h_{k,l}, \chi)$ et $p(\mathbf{x}^h_{k,j'} \mid \chi)$, respectivement, les distributions jointes prédictives a posteriori et a priori dans le bloc (k, l). Après chaque itération, le paramètre de concentration β_h est également mis à jour.

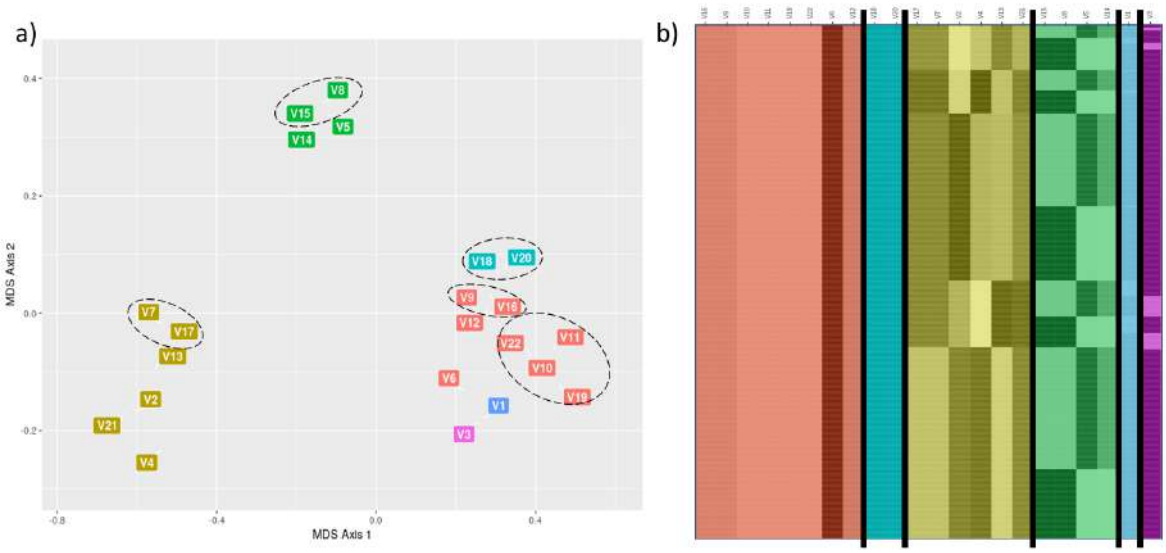


Figure 2: Résultats de l'application sur données réelles

Application sur données réelles Après validation de MCC sur des données issues du modèle génératif, nous l'appliquons sur un jeu de données réel issu de la validation d'un système d'aide au maintien dans la voie. Les groupes de variable partageant le même clustering ligne sont regroupés par couleur dans le graphique 2-a) et par distribution (pointillés). Le graphique 2-b) représente les partitions lignes. En plus de discriminer les variables non-informatives et dupliquées, ce graphique fait apparaître une structure hiérarchique entre les groupes de variables "Kaki" (qui regroupe des variables de direction), "Vert" (regroupant des variables de positions dans la voie) et l'activation des

systèmes d'aide à la conduite (V1 et V3). Les prochaines étapes sont maintenant la mise en relation de ces partitions avec les paramètres d'entrée de la simulation pour construire un modèle explicatif.

Bibliographie

Goffinet, E, Lebbah, M, Azzag, H., Giraldi, L. (2020). Autonomous Driving Validation With Model-Based Dictionary Clustering. ECML-PKDD, 2020.

Marbac, M., Vandewalle, V. (2019). A tractable multi-partitions clustering. Computational Statistics & Data Analysis, 132, 167-179.

Vandewalle, V. (2020). Multi-Partitions Subspace Clustering. Mathematics, 8(4), 597.

Hu, J., and Pei, J. (2018). Subspace multi-clustering: a review. Knowledge and information systems, 56(2), 257-284.

Guan, Y., Dy, J. G., Niu, D., and Ghahramani, Z. (2010). Variational inference for nonparametric multiple clustering. In MultiClust Workshop, KDD-2010.

Meeds, E., Roweis, S. (2007). Nonparametric bayesian biclustering. Technical report, University of Toronto.

Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., ... and Doya, K. (2017). Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions. PloS one, 12(10), e0186566.

Bouveyron, C., Bozzi, L., Jacques, J., Jollois, F. X. (2018). The functional latent block model for the co-clustering of electricity consumption curves. Journal of the Royal Statistical Society: Series C (Applied Statistics), 67(4), 897-915.

Slimen, Y. B., Allio, S., Jacques, J. (2018). Model-based co-clustering for functional data. Neurocomputing, 291, 97-108.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics, 9(2), 249-265. Meeds, E., Roweis, S. (2007). Nonparametric bayesian biclustering. Technical report, University of Toronto.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper 92-A03: Duke University.

Williamson, S., Dubey, A., Xing, E. (2013, February). Parallel Markov chain Monte Carlo for nonparametric mixture models. In International Conference on Machine Learning.

Meguelati, K., Fontez, B., Hilgert, N., Masseglia, F. (2019, April). Dirichlet process mixture models made scalable and effective by means of massive distribution. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (pp. 502-509).

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica sinica, 639-650.

Classification de données fonctionnelles multivariées par arbres binaires non-supervisées

Steven Golovkine [1] & Nicolas Klutchnikoff [2] & Valentin Patilea [3]

[1] *CREST, ENSAI, steven.golovkine@ensai.com*
[2] *IRMAR, Université Rennes 2, nicolas.klutchnikoff@univ-rennes2.fr*
[3] *CREST, ENSAI, valentin.patilea@ensai.fr*

Résumé. Nous proposons un algorithme de classification non-supervisée par modèle de mélange pour une classe générale de données fonctionnelles. Ces réalisations aléatoires peuvent être mesurées avec erreur à des points d'observations discrets, éventuellement aléatoires, dans leur domaine de définition. L'idée est de construire un ensemble d'arbres binaires par découpage récursif des observations. Le nombre de groupes est déterminé grâce aux données. Cet algorithme fournit des résultats facilement interprétables et de rapides prédictions sur de nouvelles données. Les résultats sur des données simulées montrent de bonnes performances dans différents cas complexes.

Mots-clés. Analyse en composantes principales fonctionnelles mutivariées, Classification non supervisée, Modèle de mélange

Abstract. We propose a model-based clustering algorithm for a general class of functional data. The random functional data realizations could be measured with error at discrete, and possibly random, points in the definition domain. The idea is to build a set of binary trees by recursive splitting of the observations. The number of groups are determined in a data-driven way. The algorithm provides easily interpretable results and fast predictions for online data sets. Results on simulated datasets reveal good performance in various complex settings.

Keywords. Gaussian mixtures, Model-based clustering, Multivariate functional principal components analysis

1 Introduction

Les capteurs sont de plus en plus présents dans notre vie quotidienne. Ceux-ci fournissent un grand nombre de données pouvant être modélisées comme données fonctionnelles. La quantité de données collectées de cette façon augmente rapidement, de même que leur coût d'étiquetage. Ainsi, il y a un intérêt croissant pour les méthodes qui visent à identifier des groupes homogènes au sein d'ensembles de données fonctionnelles.

Supposons un échantillon de N courbes provenant d'un même processus aléatoire, éventuellement mesurées à des instants différents et détériorées par un bruit aléatoire. Notre but est de définir une procédure, basée sur un échantillon de N courbes bruitées, permettant de constuire des groupes de courbes similaires.

2 Modèle

La structure de nos données, appellées *données fonctionnelles multivariées*, est similaire à celle présentée par Happ et Greven (2018). Les données consistent en des trajectoires indépendantes d'un processus stochastique $X = \left(X^{(1)}, \ldots, X^{(P)}\right)^\top, P \geq 1$. Pour chaque $1 \leq p \leq P$, soit $\mathcal{T}_p = [0,1]^{d_p}, d_p \geq 1$. Chaque coordonnée $X^{(p)} : \mathcal{T}_p \to \mathbb{R}$ est supposé appartenir à $\mathcal{L}^2(\mathcal{T}_p)$ muni du produit scalaire usuel, noté $\langle \cdot, \cdot \rangle_2$. Ainsi, X est un processus stochastique indexé par $\mathbf{t} = (t_1, \ldots, t_P)$ appartenant à $\mathcal{T} \coloneqq \mathcal{T}_1 \times \cdots \times \mathcal{T}_P$ et prenant ses valeurs dans $\mathcal{H} \coloneqq \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_P)$. Considérons la fonction $\langle\!\langle \cdot, \cdot \rangle\!\rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$,

$$\langle\!\langle f, g \rangle\!\rangle \coloneqq \sum_{p=1}^{P} \langle f^{(p)}, g^{(p)} \rangle_2, \quad f, g \in \mathcal{H}.$$

Happ et Greven (2018) montrent que $\mathcal{H}$ est un espace de Hilbert pour le produit scalaire $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. Soit K un entier positif, et soit Z une variable aléatoire prenant valeurs dans $\{1, \ldots, K\}$ tel que

$$\mathbb{P}(Z = k) = p_k \quad \text{avec} \quad p_k > 0 \quad \text{et} \quad \sum_{k=1}^{K} p_k = 1.$$

La variable Z représente l'appartenance à un groupe des réalisations du processus. Nous considérons que le processus stochastique X suit un *modèle de mélange fonctionnel à K composantes*, qui permet la représentation suivante :

$$X(\mathbf{t}) = \sum_{k=1}^{K} \mu_k(\mathbf{t}) \mathbf{1}_{\{Z=k\}} + \sum_{j \geq 1} \xi_j \phi_j(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T}, \tag{1}$$

où

- $\mu_1, \ldots, \mu_K \in \mathcal{H}$ sont les courbes moyennes par groupe.
- $\{\phi_j\}_{j \geq 1}$ est une base orthonormale de $\mathcal{H}$.
- Les $\xi_j, j \geq 1$ sont des variables aléatoires de $\mathbb{R}$ et conditionnellement indépendantes sachant Z. Pour chaque $1 \leq k \leq K$, $\xi_j | Z = k \sim \mathcal{N}(0, \sigma_{kj}^2)$ pour tout $j \geq 1$.

Lemme 1 *Soit X défini par le modèle* (1) *pour une certaine base orthonormale $\{\phi_j\}_{j \geq 1}$. Soit $\{\psi_j\}_{j \geq 1}$ une autre base orthonormale de $\mathcal{H}$, considérons*

$$c_j = \langle\!\langle X - \mu, \psi_j \rangle\!\rangle, \quad j \geq 1 \quad \textit{avec} \quad \mu(\cdot) = \sum_{k=1}^{K} p_k \mu_k(\cdot).$$

Alors

$$c_j | Z = k \sim \mathcal{N}(m_{kj}, \tau_{kj}^2), \quad \textit{où} \quad m_{kj} = \langle\!\langle \mu_k - \mu, \psi_j \rangle\!\rangle \quad \textit{et} \quad \tau_{kj}^2 = \sum_{l \geq 1} \langle\!\langle \phi_l, \psi_j \rangle\!\rangle^2 \sigma_{kl}^2.$$

Remarque 1 *Le lemme 1 montre que, peu importe le choix de la base orthonormée* $\{\psi_j\}_{j\geq 1}$*, les groupes seront préservés après avoir projeté les observations dans cette base. Cependant, au regard de l'objectif, certaines bases seront plus adaptées que d'autres.*

En pratique, il n'est pas possible d'utiliser un nombre infini d'élément dans la base $\{\psi_j\}_{j\geq 1}$, et la représentation (1) doit donc être tronquée. On peut montrer, sans utiliser l'hypothèse de normalité des coefficients, que cette troncature peut être arbitrairement précise. Supposons donc que la représentation (1) est tronquée à J termes. La base construite par analyse en composantes principales fonctionnelles multivariées (MFPCA) est celle qui induit la meilleure approximation (en terme de variance expliquée) à J donné (cf. Happ et Greven (2018)). Ainsi, parmi les bases utilisables en pratique, la base MFPCA sera à privilégier dans l'optique de la classification non-supervisée.

2.1 Estimation des paramètres

En pratique, les réalisations de X sont généralement mesurées avec erreur à des points d'observations discrets, éventuellement aléatoires, dans leur domaine de définition. Pour chaque $1 \leq n \leq N$, et étant donné un vecteur d'entiers positifs $\mathbf{M}_n = (M_n^{(1)}, \ldots, M_n^{(P)})$, considérons $T_{n,\mathbf{m}} = (T_{n,m_1}^{(1)}, \ldots, T_{n,m_p}^{(P)}), 1 \leq m_p \leq M_n^{(p)}, 1 \leq p \leq P$ comme étant les points d'observations aléatoires pour la courbe X_n. L'entier $M_n^{(p)}$ représente le nombre de points d'échantillonnage pour la composante p de la courbe X_n. Ces instants sont obtenus comme étant des réalisations indépendantes d'une variable aléatoire T prenant ses valeurs dans $\mathcal{T}$. Les vecteurs $\mathbf{M}_1, \ldots, \mathbf{M}_N$ représentent un échantillon indépendant d'un vecteur aléatoire $\mathbf{M}$ de moyenne $\mu_{\mathbf{M}}$ qui croit avec N. Nous supposons que les réalisations de X, $\mathbf{M}$ et T sont mutuellement indépendantes. Ainsi, nous observons les paires $(Y_{n,\mathbf{m}}, T_{n,\mathbf{m}}) \in \mathbb{R}^P \times \mathcal{T}$, où $\mathbf{m} = (m_1, \ldots, m_P), 1 \leq m_p \leq M_n^{(p)}, 1 \leq p \leq P$ avec $Y_{n,\mathbf{m}}$ défini comme

$$Y_{n,\mathbf{m}} = X_n(T_{n,\mathbf{m}}) + \varepsilon_{n,\mathbf{m}}, \quad 1 \leq n \leq N, \tag{2}$$

et les $\varepsilon_{n,\mathbf{m}}$ sont des réalisations indépendantes de $\varepsilon \in \mathbb{R}^P$ centré et de variance finie.

Les estimateurs des fonctions moyenne et covariance d'une composante $X^{(p)}, 1 \leq p \leq P$ du processus X peuvent, par exemple, être estimés en utilisant Yao, Müller et Wang (2005). Concernant l'estimation des fonctions propres et des valeurs propres de la MPFCA, ainsi que les projections des observations sur la base de fonctions propres, nous utilisons Happ et Greven (2018).

3 fCUBT

Soit $\mathcal{S}_N = \{X_1, \ldots, X_N\}$ un ensemble de réalisations du processus défini en (1). Nous considérons le problème d'apprentissage d'une partition $\mathcal{U}$ tel que chaque élément U de $\mathcal{U}$ contienne des éléments de $\mathcal{S}_N$ similaires. Notre procédure suit les idées de l'algorithme

CUBT, proposé par Fraiman, Ghattas et Svarc (2013) que l'on adapte aux données fonctionelles. Dans la suite, nous décrivons l'algorithme de clustering fonctionnel par arbres binaires non-supervisées (fCUBT).

3.1 Construction de l'arbre maximal

Dans la suite, notons $\mathfrak{T}$, un arbre binaire complet représentant une partition imbriquée de $\mathcal{S}_N$, et $\mathfrak{D} \geq 1$ sa profondeur. Soit $\mathfrak{S}_{0,0}$ le noeud racine auquel on assigne l'ensemble $\mathcal{S}_N$. Chaque noeud $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \subset \mathcal{S}_N$ est indexé par la paire $(\mathfrak{d},\mathfrak{j})$ où $0 \leq \mathfrak{d} < \mathfrak{D}$ est l'indice de profondeur et $0 \leq \mathfrak{j} < 2^{\mathfrak{d}}$ est l'indice du noeud. Chaque noeud, non terminal, $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ a deux enfants, $\mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}}$ et $\mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}+1}$, tel que

$$\mathfrak{S}_{\mathfrak{d},\mathfrak{j}} = \mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}} \cup \mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}+1}.$$

Un arbre $\mathfrak{T}$ est ainsi défini par un découpage récursif des observations. À chaque étape, un noeud $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ est potentiellement découpé en deux s'il remplit certaines conditions. Une MFPCA avec $\mathtt{n_{comp}}$ composantes, $\mathtt{n_{comp}} \leq J$, est faite sur les éléments de $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$. Il en résulte un ensemble de fonctions propres, associé à un ensemble de valeurs propres. Nous pouvons construire la matrice $C_{\mathfrak{d},\mathfrak{j}}$ dont les colonnes sont les projections des éléments de $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ sur l'ensemble des fonctions propres. Pour chaque $K = 1, \ldots, K_{max}$, nous ajustons un modèle de mélange gaussien (GMM) sur les colonnes de $C_{\mathfrak{d},\mathfrak{j}}$ par algorithme EM. Les modèles sont notés $\{\mathcal{M}_1, \ldots, \mathcal{M}_{K_{max}}\}$. Le nombre de groupes dans le noeud est estimé avec le BIC,

$$\widehat{K}_{\mathfrak{d},\mathfrak{j}} = \underset{K=1,\ldots,K_{max}}{\arg\max}\ \mathrm{BIC}(\mathcal{M}_K)$$

Si $\widehat{K}_{\mathfrak{d},\mathfrak{j}} > 1$, le noeud $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ est coupé en deux en utilisant le modèle $\mathcal{M}_2$. Sinon, ce noeud est considéré comme étant un noeud terminal, et la construction de l'arbre est stoppée pour celui-ci.

La procédure continue jusqu'à ce que l'un des critères d'arrêt soit satisfait : qu'il y ait moins de `minsize` observations dans le noeud ou alors que l'estimation $\widehat{K}_{\mathfrak{d},\mathfrak{j}}$ du nombre de clusters dans le noeud $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ soit égal à 1. Quand l'algorithme est terminé, un label est assigné à chaque feuille de l'arbre.

3.2 Étape de jointure

En théorie, l'arbre a le même nombre de feuilles que le processus X a de composantes. En pratique, c'est rarement le cas et le nombre de feuilles peut être bien supérieur au vrai nombre de groupes. Ainsi, une étape de jointure, dont l'idée est d'associer des noeuds qui n'ont pas le même ascendant, peut être nécessaire.

Soit $\mathcal{G} = (V, E)$ un graphe où $V = \{\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, 0 \leq j \leq 2^{\mathfrak{d}}, 0 \leq \mathfrak{d} < \mathfrak{D} | \mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \text{ est une feuille}\}$ est l'ensemble des sommets et

$$E = \left\{(\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'}) | \mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'} \in V, \mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \neq \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'} \text{ et } \widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')} = 1\right\}$$

est l'ensemble des arêtes. $\widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')}$ est l'estimation du nombre de groupes dans $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}\cup\mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'}$ en utilisant la même méthodologie que pour l'étape précédente.

Pour chaque élément $(\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'})$ de E, nous associons la valeur du BIC qui correspond à $\widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')}$. L'arête de $\mathcal{G}$ ayant la plus grande valeur du BIC est ensuite supprimée, et les sommets associés sont joints. Il y a donc un groupe en moins. Cette procédure est lancée récursivement jusqu'à ce qu'aucun noeud ne puisse être joint avec un autre ou bien qu'il n'y ait plus qu'un noeud dans l'arbre.

Une fois la partition $\mathcal{U}$ créée, nous pouvons utiliser celle-ci pour classifier de nouvelles observations. Cette classification se fait par descente de l'arbre $\mathfrak{T}$, et nous pouvons donc calculer les probabilités d'appartenance à chacune des classes pour les nouvelles observations.

4 Analysis empirique

Montrons les performances de notre algorithme sur un exemple et fixons $K = 5$, $P = 2$, $\mathcal{T}_1 = \mathcal{T}_2 = [0, 1]$. Un échantillon de $N = 1000$ courbes bivariées indépendantes est simulé suivant le modèle : pour $t_1, t_2 \in [0, 1]$,

$$
\begin{array}{lll}
\text{Cluster 1:} & X^{(1)}(t_1) = h_1(t_1) + b_{0.9}(t_1), & \\
& X^{(2)}(t_2) = h_3(t_2) + 1.5 \times b_{0.8}(t_2) & \\
\text{Cluster 2:} & X^{(1)}(t_1) = h_2(t_1) + b_{0.9}(t_1), & \\
& X^{(2)}(t_2) = h_3(t_2) + 0.8 \times b_{0.8}(t_2), & \\
\text{Cluster 3:} & X^{(1)}(t_1) = h_1(t_1) + b_{0.9}(t_1), & \\
& X^{(2)}(t_2) = h_3(t_2) + 0.2 \times b_{0.8}(t_2), & \\
\text{Cluster 4:} & X^{(1)}(t_1) = h_2(t_1) + 0.1 \times b_{0.9}(t_1), & \\
& X^{(2)}(t_2) = h_2(t_2) + 0.2 \times b_{0.8}(t_2), & \\
\text{Cluster 5:} & X^{(1)}(t_1) = h_3(t_1) + b_{0.9}(t_1), & \\
& X^{(2)}(t_2) = h_1(t_2) + 0.2 \times b_{0.8}(t_2). &
\end{array}
$$

Les fonctions h sont définies, par $h_1(t) = (6 - |20t - 6|)_+ /4$, $h_2(t) = (6 - |20t - 14|)_+ /4$ et $h_3(t) = (6 - |20t - 10|)_+ /4$, pour $t \in [0, 1]$. Les fonctions b_H sont définies, pour $t \in [0, 1]$, par $b_H(t) = (1+t)^{-H} B_H(1+t)$ avec $B_H(\cdot)$ est un mouvement brownien fractionnaire avec un coefficient de Hurst H. Les proportions du mélange sont supposées égales.

Les données sont obtenues suivant le modèle (2). Chaque courbe est observée à 101 points répartis aléatoirement sur $[0, 1]$. Les vecteurs d'erreurs sont supposés être de lois normales centrées et de variance $1/2$. Pour chaque $n \in \{1, \ldots, N\}$, nous observons une réalisation du vecteur $X = \left(X^{(1)} + \alpha X^{(2)}, X^{(2)}\right)^\top$, avec $\alpha = 0.4$.

Notre méthode est comparée aux algorithmes `FunHHDC` (Schmutz et al. (2020)), `Funclust` (Jacques et Preda (2014)) et k-`means` (Ieva et al. (2013)) sur les courbes (k-`means`-d_1) et leur dérivées (k-`means`-d_2)). Nous évaluons aussi notre performance par rapport à un GMM après une FPCA (`FPCA+GMM`) et à notre algorithme si on ne fait que la construction de l'arbre sans l'étape de jointure (`Growing`). Notre algorithme montre de bonnes performances en terme d'index de Rand (cf. Table 1) et d'estimation du nombre de groupes dans le jeu de données (cf. Figure 1).

Method	1	2	3	4	5	6	7+
`fCUBT`	-	-	-	-	0.664	0.238	0.098
`Growing`	-	-	-	-	0.604	0.182	0.214
`FPCA+GMM`	-	-	-	-	0.414	0.396	0.19
`FunHDDC`	0.508	0.492	-	-	-	-	-
`Funclust`	-	0.066	0.182	0.192	0.200	0.196	0.164
k-means-d_1	-	-	-	-	0.034	0.144	0.822
k-means-d_2	-	0.004	0.01	0.094	0.874	0.010	0.008

Table 1: Nombres de groupes

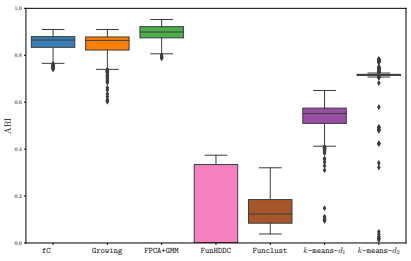


Figure 1: Index de Rand

Enfin, la méthodologie est dévelopée pour des données fonctionnelles multivariées pouvant être définies sur des domaines différents et potentiellement de différentes dimensions. Ainsi, nous pouvons considérer des processus définis sur un carré dans le plan, $\mathcal{T} = [0, 1]^2$ par exemple. Dans ce cas, une décomposition de ce processus peut être faite, par exemple, grâce à l'algorithme FCP-TPA pour une décomposition tensorielle (Allen (2013)) et donc être utilisé dans la MFPCA. Une version complète de l'article est disponible à l'adresse suivante : arxiv:2012.05973.

Bibliographie

Allen, G. I. (2013). Multi-way functional principal components analysis. In 2013 *5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 220-223.

Fraiman, R., Ghattas, B. and Svarc, M. (2013). Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7.

Happ, C. and Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113 649-659.

F. Ieva, A. M. Paganoni, D. Pigoli, and V. Vitelli. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society.* Series C (Applied Statistics), 62(3):401-418, 2013.

Jacques, J. and Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71 92-106.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L. and Martin, P. (2020). *Clustering multivariate functional data in group-specific functional subspaces.* Computational Statistics

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100 577-590.

Incrémentation séquentielle de la dimension en apprentissage statistique

Thierry Gonon [1] & Christophette Blanchet-Scalliet [2] & Céline Helbert [3] & Bruno Demory [4]

[1] *Univ Lyon, Ecole Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan, 36 avenue Guy de Collongue, F-69134 Ecully Cedex, France, thierry.gonon@ec-lyon.fr*

[2] *Univ Lyon, Ecole Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan, 36 avenue Guy de Collongue, F-69134 Ecully Cedex, France, christophette.blanchet@ec-lyon.fr*

[3] *Univ Lyon, Ecole Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan, 36 avenue Guy de Collongue, F-69134 Ecully Cedex, France, celine.helbert@ec-lyon.fr*

[4] *Valeo Systèmes Thermiques, 8 Rue Louis Lormand, 78321 La Verrière, France, bruno.demory@valeo.com*

Résumé. Avant d'être mis sur le marché, les produits industriels font l'objet d'une batterie de tests dont une partie sont des simulations numériques. Les codes numériques utilisés modélisent l'évolution de variables d'intérêt (sorties du code) d'ordre physique, en fonction de variables d'entrée géométriques ou environnementales. Ces simulations sont complexes car elles font intervenir beaucoup de variables d'entrée, donc l'espace d'entrée du modèle est de grande dimension. Des études sont réalisées pour comprendre l'influence des variables d'entrée sur les variables de sortie. Elles utilisent des métamodèles qui sont des modèles simples donnant des prédictions rapides approchant une des variables de sortie. Les métamodèles sont entraînés sur un plan d'expérience : un ensemble de points de l'espace d'entrée (qui correspondent à des valeurs des variables d'entrées) pour lesquels on connaît déjà la valeur de la variable de sortie. Les premières études se font en petite dimension, se focalisant sur un ensemble restreint de variables d'entrée, les laissant libre de varier pour regarder leur influence sur la variable de sortie, les autres variables d'entrée étant fixées à des valeurs nominales. Puis les études se complexifient en augmentant progressivement la dimension. La manière classique de traiter cette augmentation de la dimension est de recommencer à zéro chaque fois que de nouvelles variables d'entrée sont ajoutées, en régénérant un nouveau plan d'expérience à simuler, indépendant des plans précédents, et en entraînant un nouveau métamodèle sur ce plan. Cette approche est gourmande en temps de calcul et entraîne une perte d'information car les plans d'expérience précédents ne sont pas réutilisés. Une approche alternative, proposée dans notre travail, est de mettre à jour progressivement le plan d'expérience et le métamodèle, en tenant compte des précédents. La métamodélisation utilise la régression par Processus Gaussien. A chaque étape k, la variable de sortie est supposée être la réalisation d'un Processus Gaussien qui est la somme de deux processus : le premier est défini sur l'espace d'entrée de l'étape $k-1$ (qui est un sous-espace de l'espace d'entrée de l'étape k). Le second

est construit indépendamment sur l'espace d'entrée de l'étape k mais est nul sur le sous-espace de l'étape $k-1$. Notre travail a consisté en la recherche de processus candidats s'annulant sur un sous-espace et en l'implémentation de l'algorithme EM (Expectation-Maximization) pour estimer les paramètres du processus complet.

Mots-clés. Métamodèle, Krigeage, Régression par Processus Gaussien, Grande dimension, Conditionnement infini, Multifidélité ...

Abstract. Industrial products are studied numerically before being sold. The corresponding numerical codes involve geometrical or environmental input variables and physical output variables. They are complex as they involve lots of input variables, so the dimension of the input space is high. Studies are provided to quantify the influence of the input variables on the behavior of the output variables. They use metamodels which are simpler models that give quick predictions approaching the true values of one of the output variables. The metamodels are trained on a Design of Experiments (DoE) : a set of points of the input space (corresponding to values of the input variables) for which the value of the output variable is already known. The first studies are done in small dimension, focusing on a small amount of important input variables, letting them vary freely to see their influence on the output variable, while the other input variables are fixed to nominal values. Then, the studies are complexified by progressively increasing the dimension. The classical way of treating the increasing dimension is to start from scratch each time new input variables are released, regenerating a DOE and a metamodel independent from the previous ones. This approach can be time consuming and the previous data is lost because unused. An alternative, presented in our work, is to upgrade gradually the design and the metamodel, based on the previous ones. This enables to exploit all the available data. The surrogate approach is based on Gaussian Process regression. At each step k, the output variable is supposed to be the realization of a Gaussian Process which is the sum of two processes : the first one is defined on the previous input space of step $k-1$ (subspace of the current input space of step k), the second is built independently on the input space of step k but is conditionned to be null on the subspace of step $k-1$. Our work has consisted in searching candidate processes that are null on a subspace and implementing the EM (Expectation-Maximization) algorithm to estimate the parameters of the complete process.

Keywords. Metamodel, Kriging, Gaussian Process Regression, Big dimension, Infinite conditionning, Multifidelity ...

1 Cas d'étude

On considère l'exemple d'une sortie y fonction de quatre entrées $x = (x_1, x_2, x_3, x_4)$: $y(x_1, x_2, x_3, x_4) = f_1(x_1, x_2) + f_2(x_1, x_2, x_3, x_4)$, avec

$$\begin{cases} f_1(x_1, x_2) & = & \left[4 - 2.1(4x_1 - 2)^2 + \frac{(4x_1-2)^4}{3}\right](4x_1 - 2)^2 \\ & + & (4x_1 - 2)(2x_2 - 1) + \left[-4 + 4(2x_2 - 1)^2\right](2x_2 - 1)^2 \\ f_2(x_1, x_2, x_3, x_4) & = & 4\exp\left(-\|x - 0.3\|^2\right) \end{cases}$$

On considère que l'on réalise deux études successives :

- Lors de l'étude 0, on se focalise sur $y(x_1, x_2, \frac{x_1+x_2}{2}, 0.2x_1 + 0.7)$. On laisse varier les deux premières entrées (x_1, x_2) et on impose les valeurs des deux suivantes : $x_3 = \frac{x_1+x_2}{2}$, $x_4 = 0.2x_1 + 0.7$. On dispose pour cette étude d'un plan d'expérience sur lequel on construit un métamodèle.
- Lors de l'étude 1, on regarde $y(x_1, x_2, x_3, x_4)$. Les deux dernières entrées (x_3, x_4) sont libérées. On dispose d'un nouveau plan d'expérience pour cette étude. Le but est de construire un métamodèle à partir de toute l'information (tous les plans d'expérience) dont on dispose. Une modélisation simple serait de proposer une régression par Processus Gaussien conditionné à l'intégralité de l'information (tous les plans d'expérience). On propose une modélisation différente qui tient compte de la séquentialité des études et donc de l'obtention des observations.

2 Modélisation

La sortie y est modélisée par un Processus Gaussien Y dont on considère qu'elle est une réalisation. En nous inspirant du modèle de la Multifidélité de Kennedy et O'Hagan (2000), on choisit de décomposer Y comme suit :

$$Y(x_1, x_2, x_3, x_4) = \begin{cases} Y_0(x_1, x_2) & = & Z_0(x_1, x_2) & \text{si} & \left| \begin{array}{lcl} x_3 & = & \frac{x_1+x_2}{2} \\ x_4 & = & 0.2x_1 + 0.7 \end{array} \right. \\ Y_1(x_1, x_2, x_3, x_4) & = & Y_0(x_1, x_2) + Z_1(x_1, x_2, x_3, x_4) & \text{sinon} \end{cases}$$

Z_0 et Z_1 sont des Processus Gaussiens indépendants et Z_1 s'annule sur une partie de l'espace des entrées :

$$Z_1(x_1, x_2, \tfrac{x_1+x_2}{2}, 0.2x_1 + 0.7) = 0 \quad \forall x_1, x_2$$

On fait face à deux difficultés : la construction d'un processus Z_1 s'annulant sur un continuum, et l'estimation jointe des paramètres des distributions de Z_0 et Z_1 à partir des plans d'expérience à disposition.

3 Candidats pour Z_1

On propose deux candidats pour Z_1. Il s'agit de Processus Gaussiens non stationnaires mais construits à partir d'un Processus Gaussien centré stationnaire d'ordre 2 (covariance stationnaire) noté $\widetilde{Z}_1 \sim \mathcal{GP}(0, \widetilde{k}(x, x'))$:

- $Z_1(x_1, x_2, x_3, x_4) = \widetilde{Z}_1(x_1, x_2, x_3, x_4) - \widetilde{Z}_1(x_1, x_2, \frac{x_1+x_2}{2}, 0.2x_1 + 0.7)$ que l'on appelle Red (pour réduit). C'est un Processus Gaussien centré de noyau de covariance égal à :

$$\begin{aligned} k(x, x') &= \widetilde{k}(x, x') + \widetilde{k}\left((x_1, x_2, \tfrac{x_1+x_2}{2}, 0.2x_1 + 0.7), (x'_1, x'_2, \tfrac{x'_1+x'_2}{2}, 0.2x'_1 + 0.7)\right) \\ &- \widetilde{k}\left(x, (x'_1, x'_2, \tfrac{x'_1+x'_2}{2}, 0.2x'_1 + 0.7)\right) - \widetilde{k}\left((x_1, x_2, \tfrac{x_1+x_2}{2}, 0.2x_1 + 0.7), x'\right) \end{aligned}$$

- $Z_1(x_1, x_2, x_3, x_4) = \left[\widetilde{Z}_1(x_1, x_2, x_3, x_4) \mid \widetilde{Z}_1(t_1, t_2, \frac{t_1+t_2}{2}, 0.2t_1 + 0.7) = 0 \forall t_1\right]$ (c'est une généralisation de la notation de conditionnement sur un ensemble infini continu de points). Ce Processus est décrit dans le travail de Gauthier (2011). On l'appelle P (pour préconditionné). C'est un Processus Gaussien centré de noyau de covariance égal à :

$$k(x, x') = \widetilde{k}(x, x') - \sum \lambda_n \phi_n(x) \phi_n(x')$$

avec les ϕ_n définies par

$$\phi_n(x) = \frac{1}{\lambda_n} \int \widetilde{k}\left(x, (t_1, t_2, \frac{t_1 + t_2}{2}, 0.2t_1 + 0.7)\right) \widetilde{\phi}_n(t_1, t_2) dt_1 dt_2$$

and les $(\lambda_n, \widetilde{\phi}_n)$ solutions du problème aux valeurs propres suivant :

$$\int \widetilde{k}\left((x_1, x_2, \frac{x_1 + x_2}{2}, 0.2x_1 + 0.7), (t_1, t_2, \frac{t_1 + t_2}{2}, 0.2t_1 + 0.7)\right) \widetilde{\phi}_n(t_1, t_2) dt_1 dt_2 = \lambda_n \widetilde{\phi}_n(x_1, x_2)$$

4 Estimation des paramètres de Y

Étant donnés les plans d'expérience des études 0 et 1 : $(\mathbb{X}_0, \mathbb{Y}_0)$ et $(\mathbb{X}_1, \mathbb{Y}_1)$, on cherche à estimer les paramètres de Z_0 et Z_1 qui sont respectivement (avec dans l'ordre le paramètre de moyenne, le paramètre de variance et les paramètres de covariance) $\eta_0 = (m_0, \sigma_0^2, \theta_0 = (\theta_{01}, \theta_{02}))$ et $\eta_1 = (m_1, \sigma_1^2, \theta_1 = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}))$. Pour cela on maximise la vraisemblance du modèle complet $\mathcal{L}(\eta_0, \eta_1; \mathbb{Y}_0, \mathbb{Y}_1)$ qui est en grande dimension (10 paramètres à optimiser). On utilise l'algorithme d'Expectation-Maximization (décrit par Friedman, Hastie

et Tibshirani (2001)). Il s'agit d'une méthode itérative : on part d'un jeu de paramètres initial, puis le jeu de paramètres est mis à jour bout par bout à partir du jeu précédent.

$$\left\{ \begin{array}{rcl} \eta_0^{(i)} & = & argmin_{\eta_0} \mathcal{Q}_0(\eta_0, \eta^{(i-1)}) \\ \eta_1^{(i)} & = & argmin_{\eta_1} \mathcal{Q}_1(\eta_1, \eta^{(i-1)}) \end{array} \right.$$

$\mathcal{Q}_0$ et $\mathcal{Q}_1$ sont les espérances des vraisemblances de Z_0 et Z_1 (qui ne sont pas entièrement connues) conditionnées par les plans d'expériences : $\mathbb{E}[\mathcal{L}(\eta_k^{(i)}; \mathbb{Z}_k) \mid y(\mathbb{X}_0) = \mathbb{Y}_0, y(\mathbb{X}_1) = \mathbb{Y}_1]$ (où $\mathbb{Z}_k$ est l'observation de Z_k qui n'est pas connue). On recherche notamment un compromis entre parcimonie des paramètres de Z_1 (pour la robustesse de leur estimation) et flexibilité (pour l'ajustement aux données d'entraînement).

5 Résultats

On compare notre méthodologie à un métamodèle classique de krigeage (voir figure 1). L'information des études précédentes est utile car elle permet de faire décroître l'erreur de prédiction du krigeage. D'autre part, pour cet exemple, la méthode avec Red permet un meilleur résultat que le krigeage.

Bibliographie

Kennedy, M. C. et O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available, *Biometrika*, 87, pp. 1-13.
Gauthier, B. (2011). *Approche spectrale pour l'interpolation à noyaux et positivité conditionnelle*, Ecole Nationale Supérieure des Mines de Saint-Etienne.
Friedman, J., Hastie, T. et Tibshirani, R. (2001). *The elements of statistical learning*, Springer series in Statistics, New York.

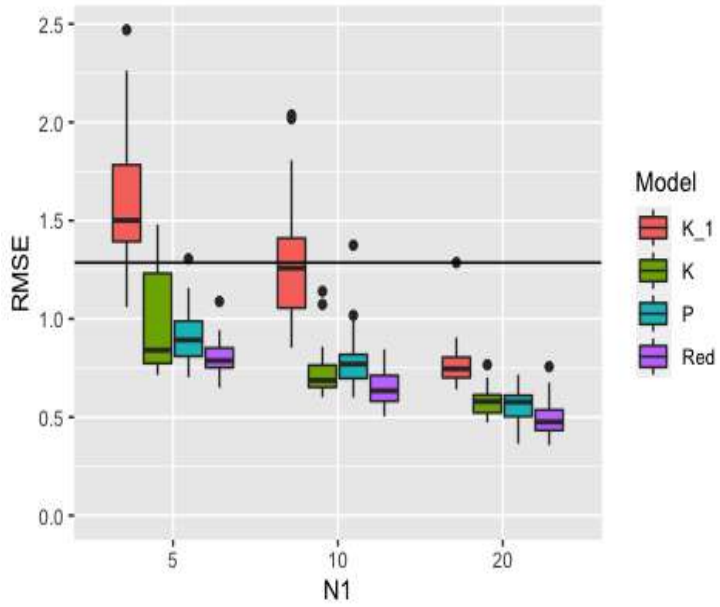


Figure 1: RMSE des différents modèles pour la fonction étudiée pour différentes tailles du plan d'entraînement de l'étude 1 (N1), la taille du plan d'entraînement de l'étude 0 étant fixé à 10. K_1 est un krigeage entraîné uniquement sur le plan d'expérience de l'étude 1, K est un krigeage entraîné sur les plans d'expérience des études 0 et 1. P est la méthode utilisant le processus P, Red est la méthode utilisant le processus Red.

Regularity of the center-outward transport based distributions and quantile functions.

Alberto González-Sanz

Institut de Mathématiques de Toulouse
alberto.gonzalez_sanz@math.univ-toulouse.fr

Abstract. We provide sufficient conditions under which the center-outward distribution and quantile functions introduced in Chernozhukov et al. (2017) and Hallin (2017) are homeomorphisms, thereby extending a recent result by Figalli (2017). Finally we derive some properties of this extension of the distribute functions. All these results can be found in del Barrio et al. (2020).

Keywords. Monge–Ampère equation, Multivariate ranks, Optimal transportation, Quantile contours.

Let P be a Borel probability measure on the real line with continuous distribution function F and denote by $\mathrm{U}_{[0,1]}$ the uniform distribution on $(0,1)$: then F is the unique gradient of a convex function such that $T\sharp \mathrm{P} = \mathrm{U}_{[0,1]}$, where $T\sharp P$ denotes the *push forward of* P *by* T—namely, the distribution of $T(X)$ under $X \sim \mathrm{P}$ (T being a measurable map from $\mathbb{R}$ to $(0,1)$). With generalization to higher dimension in mind, however, Chernozhukov et al. (2018) and Hallin (2017) rather consider $\mathbf{F}_{\pm} = 2F(x) - 1$, the so-called *center-outward distribution function* of P, being the unique gradient of a convex function such that $T\sharp \mathrm{P} = \mathrm{U}_{[0,1]}$, where U_1 is the uniform distribution over $(-1,1)$, the one-dimensional unit ball $\mathbb{B}_1$.

The latter definition, indeed, readily extends to arbitrary dimensions. Let P denote a Borel probability measure on $\mathbb{R}^d$ with Lebesgue density p. McCann (1995) proved that there exists a P-a.s. unique map $\mathbf{F}_{\pm}$ coinciding P-a.s. with the Lebesgue-a.e. gradient $\nabla\varphi$ of a convex function φ and $\mathbf{F}_{\pm}\sharp \mathrm{P} = \mathrm{U}_d$, where U_d denotes the *uniform* distribution over the open d-dimensional unit ball $\mathbb{B}_d$, that is, U_d here corresponds to the uniform choice of a direction on the unit sphere $\mathbb{S}_{d-1} := \bar{\mathbb{B}}_d - \mathbb{B}_d$ in $\mathbb{R}^d$ combined with an independent uniform choice in $(0,1)$ of a distance to the origin.

This definition of the center-outward distribution suffers from one main limitation, the distribution function $\mathbf{F}_{\pm}$ is only defined P-a.s.; this means, for instance, that $\mathbf{F}_{\pm}$ is not well defined outside the support of P. This limitation, however, disappears if P is such that φ is everywhere differentiable. That this is indeed the case was shown by Figalli (2018) for P in the class of distributions P with densities p and support $\mathcal{X} = \mathbb{R}^d$ satisfying Assumption A below. For any P in that class of distributions, Figalli actually establishes that $\nabla\varphi(\mathbf{x})$ is well defined for all $\mathbf{x}$ and, when restricted to

$$\mathbb{R}^d_{(\mathbf{0})} := \mathbb{R}^d \setminus \{\mathbf{x} : \nabla\varphi(\mathbf{x}) = \mathbf{0}\},$$

it defines a homeomorphism between $\mathbb{R}^d_{(\mathbf{0})}$ and the punctured ball $\mathbb{B}_d \setminus \{\mathbf{0}\}$.

The main theorem provides simple sufficient conditions for Figalli's results to hold in a wider class beyond the following assumption.

Assumption A. For every $R > 0$, there exist constants $0 < \lambda_R \leq \Lambda_R$ such that

$$\lambda_R \leq p(\mathbf{x}) \leq \Lambda_R \quad \text{for all } \mathbf{x} \in \mathcal{X} \cap R\,\mathbb{B}_d. \tag{1}$$

Moreover, considering the inverse problem—the regularity of the transportation map from U_d to P, defined as the unique a.s. defined map $\mathbf{Q}_\pm$ being the gradient of the convex conjugate of φ, namely $\psi = \varphi^*$—Theorem 1 establishes sufficient conditions under which $\mathbf{Q}_\pm$ and $\mathbf{F}_\pm$ are inverses of each other when restricted to $\mathbb{B}_d \setminus \{\mathbf{0}\}$ and $\mathcal{X} \setminus K$, for some compact convex set K with Lebesgue measure 0.

Beyond other theoretical considerations, these are the key properties required to prove a.s. convergence of the empirical center-outward distribution functions to their theoretical counterparts (see del Barrio et al. (2018)). Hence, the following result also is extending the validity of the center-outward Glivenko-Cantelli theorem in that reference.

Theorem 1 *Let* P *be a probability measure with density p supported on the open convex set $\mathcal{X} \subseteq \mathbb{R}^d$. (i) If p satisfies* (1)*, there exists a compact convex set K with Lebesgue measure* 0 *such that the center-outward quantile function $\mathbf{Q}_\pm := \nabla\psi$ and the center-outward distribution function $\mathbf{F}_\pm := \nabla\psi^*$ are homeomorphisms between $\mathbb{B}_d \setminus \{\mathbf{0}\}$ and $\mathcal{X} \setminus K$, inverses of each other.*

(ii) If, moreover, $p \in \mathcal{C}^{k,\alpha}_{\mathrm{loc}}(\mathcal{X})$ for some $k \in \mathbb{N}$ and $\alpha \in (0,1)$, then $\mathbf{Q}_\pm$ and $\mathbf{F}_\pm$ are diffeomorphisms of class $\mathcal{C}^{k+1,\alpha}_{\mathrm{loc}}$ between $\mathbb{B}_d \setminus \{\mathbf{0}\}$ and $\mathcal{X} \setminus K$.

The proof of such a result uses the classic approach of Caffarelli's theory on the Monge-Ampére equation. We refer to Figalli (2017) and reference therein. Given an open set $\mathcal{X} \subseteq \mathbb{R}^d$ and a (finite) convex function $\varphi : \mathcal{X} \to \mathbb{R}$, denoting by ℓ_d the Lebesgue measure on $\mathbb{R}^d$, the Monge-Ampère measure associated with φ is defined by

$$\mu_\varphi(E) := \ell_d\Big(\partial\varphi(E)\Big)$$

for every Borel set $E \subseteq \mathbb{R}^d$. It can be checked that μ_φ is indeed a locally finite Borel measure on $\mathcal{X}$. The classic assumption on the Monge-Ampère measure is that it is bounded in the sense

$$\lambda\ell_d(A) \leq \mu_\psi(A) \leq \Lambda\ell_d(A), \tag{2}$$

for some constants $0 < \lambda < \Lambda$. In our case, instead of (2), we have that for every compact subset M of $\mathbb{B}_d$, there exist constants α_M and A_M such that, for every Borel set $A \subseteq M$,

$$\alpha_M\ell_d(A) \leq \mu_\psi(A) \leq A_M(\ell_d(A))^{1/d},$$

where ψ is defined in Theorem 1. del Barrio et al. (2020) proved that the latest bounds still allow to apply the usual Alexandrov estimates to show the regularity of ψ.

To conclude this note, we present some results that more or less directly follow as consequences of Theorem 1. The first one is about the asymptotic invariance of center-outward distribution functions; the second is a result on the shape of quantile contours, which turn out to satisfy a kind of relaxed version of convexity.

A classical univariate distribution function F trivially satisfies

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1,$$

hence, in terms of the univariate center-outward distribution function $F_\pm := 2F - 1$,

$$\lim_{t\to\infty} F_\pm(tu) = u \qquad \text{for all } u \text{ such that } |u| = 1.$$

Let us show that this carries over to $\mathbf{F}_\pm$ in general dimension. Keeping the notation from the previous sections, and setting a measure P with density on $\mathbb{R}^d$. For any $\mathbf{u}$ on the unit sphere $\mathbb{S}_{d-1}$, any sequence $(t_n)_{n\in\mathbb{N}}$ of real numbers such that $t_n \to \infty$, and any $\mathbf{y}_n \in \partial\varphi(t_n\mathbf{u})$,

$$\lim_{n\to\infty} \mathbf{y}_n = \mathbf{u}.$$

Then with the assumptions of Theorem 1 we have that

$$\lim_{n\to\infty} \mathbf{F}_\pm(t_n\mathbf{u}) = \mathbf{u}.$$

With regards to the shape of the center-outward quantile contours if P satisfies the assumptions of Theorem 1, then , for all $r \in (0,1)$ and all $\mathbf{y}$ belonging to the boundary of $\mathbf{Q}_\pm(r\,\mathbb{B}_d)$, there exists a ray T emanating from $\mathbf{y}$ for which $\overline{\mathbf{Q}_\pm(r\,\mathbb{B}_d)} \cap T = \{\mathbf{y}\}$.

Bibliography

del Barrio, E., Beirlant, J., Buitendag, S. and Hallin, M. (2019). Center-outward quantiles and the measurement of multivariate risk. https://arxiv.org/abs/1912.04924

del Barrio, E., Cuesta Albertos, J., Hallin, M., and Matrán, C. (2018). Smooth cyclically monotone interpolation and empirical center-outward distribution functions.

del Barrio, E. and González-Sanz, A. and Hallin, M. (2020).A note on the regularity of optimal-transport-based center-outward distribution and quantile functions.*Journal of Multivariate Analysis*, 180, pp. 104671.

Figalli, A. *The Monge-Ampère Equation and Its Applications*, Zurich Lectures in Advanced Mathematics, European Mathematical Society (EMS), Zurich, 2017.

Figalli, A. (2018). On the continuity of center-outward distribution and quantile functions, *Nonlinear Analysis*, **177**, 413–421.

Hallin, M. (2017). Distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. Journal*, **80**, 309–323.

Tests d'Homogénéité basés sur la distance de Wasserstein pour l'étude des Protéines Intrinsèquement Désordonnées

Javier González Delgado [1], Alberto González-Sanz [2],
Pierre Neuvial [3] & Juan Cortés [4]

[1] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, LAAS-CNRS, javier.gonzalez-delgado@math.univ-toulouse.fr*
[2] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, alberto.gonzalez_sanz@math.univ-toulouse.fr*
[3] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, pierre.neuvial@math.univ-toulouse.fr*
[4] *LAAS-CNRS, Université de Toulouse, CNRS, juan.cortes@laas.fr*

Résumé. L'étude structurale des Protéines Intrinsèquement Désordonnées (IDPs) requiert des méthodes statistiques afin de prendre en compte leur grande variabilité conformationnelle. Leur conformation locale est définie par une mesure sur le tore plat $\mathbb{R}^2/\mathbb{Z}^2$, et donc leur analyse repose sur l'espace $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$. La distance de Wasserstein est une métrique entre distributions naturelle pour ce problème, du fait de son adaptabilité à la géométrie de l'espace. Cette métrique permettra la quantification des modifications de la structure locale dues aux changements de la séquence, et donc à une meilleure compréhension des relations séquence-structure-fonction dans les IDPs. Dans ce travail, nous commençons par étendre certains résultats de la Théorie de Transport Optimal à l'espace $\mathbb{R}^2/\mathbb{Z}^2$, en particulier un Théorème Central Limite. Nous étudions ensuite la construction de tests d'homogénéité de deux lois sur le tore plat basés sur la distance de Wasserstein. Les différentes approches envisagées ont été évaluées à l'aide de données réelles de conformation de protéines.

Mots-clés. Transport Optimal, Distance de Wasserstein, Théorème Central Limite, Test d'homogénéité, Protéines Intrinsèquement Désordonnées.

Abstract. Due to their high conformational variability, the structural investigation of Intrinsically Disordered Proteins (IDPs) requires statistical methods. Their local conformation is defined by a measure on the flat torus $\mathbb{R}^2/\mathbb{Z}^2$, and therefore their analysis leads on the space $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$. Due to its adaptability to the space geometry, Wasserstein distance will be used as a metric between distributions. This metric will enable the quantification of local structural modifications due to changes in the protein sequence, and therefore help to better understand sequence-structure-function relationships in IDPs. This analysis starts by extending results from Optimal Transport Theory to the space $\mathbb{R}^2/\mathbb{Z}^2$, in particular a Central Limit Theorem, and continues by assessing different techniques of two-sample goodness-of-fit testing for measures on the flat torus based on Wasserstein

distance. The envisaged approaches have been evaluated via their implementation on real protein conformation data.

Keywords. Optimal Transport, Wasserstein Distance, Central Limit Theorem, Goodness-of-fit test, Intrinsically Disordered Proteins.

1 L'étude des Protéines Intrinsèquement Désordonnées

Les protéines intrinsèquement désordonnées (IDPs), à différence des protéines plus communes dites globulaires, ne se replient pas dans une forme tridimensionnelle stable et bien définie. En effet, leur séquence en acides aminés ne détermine pas un seul état conformationnel mais un grand ensemble d'états que la protéine peut adopter en solution. Cette diversité structurale représente une énorme difficulté pour leur étude avec des techniques expérimentales et/ou calculatoires. Cependant, la possibilité d'établir des liens entre séquence et propriétés structurales est essentielle pour comprendre leur fonctionnement et leurs rôles à l'intérieur de la cellule. Il est important de mentionner que les IDPs constituent environ le 40% du génome humain, et leur dysfonctionnement peut induire des maladies sévères comme le cancer ou des maladies neurodégénératives.

L'état conformationnel d'une protéine peut être défini par un vecteur d'angles, correspondants aux torsions autour des liaisons chimiques entre les atomes qui constituent son « squelette ». Ce vecteur contient deux valeurs par acide aminé, les angles ϕ et ψ, qui définissent l'état au niveau local, et donc les distributions correspondantes doivent être définies comme des mesures sur le tore plat $\mathbb{R}^2/\mathbb{Z}^2$. Des exemples sont illustrés sur la Figure (1). Par conséquent, au niveau local, l'analyse des conformations des IDPs se place sur l'espace $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$, qu'on cherche à équiper d'une métrique. La définition d'une telle distance entre les conformations des protéines est essentielle afin de mesurer des similarités parmi les structures locales, ainsi que pour construire des méthodes de description de la distribution des structures au niveau global à partir de l'information à niveau local. Ce dernier point est l'un des enjeux principaux de la Biologie Structurale actuellement.

La géométrie du problème est un aspect fondamental, qui n'a cependant pas encore été bien pris en compte dans les travaux précédents en Biologie Structurale. Avoir une distance permettant de travailler sur le vrai espace du problème est primordial. Pour cette raison, une approche basée sur la distance de Wasserstein, qui repose sur la géométrie de l'espace, semble être une bonne option. L'objectif du présent travail consiste à étendre les résultats principaux de la théorie de Transport Optimal à l'espace $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$, notamment un Théorème Central Limite, et, dans un deuxième temps, à étudier la construction de tests d'homogénéité pour deux mesures sur $\mathbb{R}^2/\mathbb{Z}^2$ basés sur la distance de Wasserstein.

2 Distances de Wasserstein sur $\mathbb{R}^2/\mathbb{Z}^2$

L'espace $\mathbb{R}^2/\mathbb{Z}^2$ peut être équipé d'une distance dérivée de la norme euclidienne $||\cdot||$,

$$d(\bar{x},\bar{z}) := \inf_{p\in\mathbb{Z}} ||x - z + p||,$$

où $\bar{x}$ désigne la classe d'équivalence de $\mathbf{x} \in \mathbb{R}^d$. Pour deux mesures de probabilité $P, Q \in \mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$, on dit qu'une mesure de probabilité $\pi \in \mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2 \times \mathbb{R}^2/\mathbb{Z}^2)$ est un *plan de transport optimal pour le coût* d^2 entre P et Q si elle est une solution de

$$\mathcal{T}_2(P,Q) := \inf_{\gamma\in\Pi(P,Q)} \int_{\mathbb{R}^2/\mathbb{Z}^2\times\mathbb{R}^2/\mathbb{Z}^2} d^2(\bar{x},\bar{z}) d\pi(\bar{x},\bar{z}), \tag{1}$$

où $\Pi(P,Q)$ est l'ensemble des mesures de probabilité $\pi \in \mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2 \times \mathbb{R}^2/\mathbb{Z}^2)$ telles que $\pi(A\times\mathbb{R}^2/\mathbb{Z}^2) = P(A)$ et $\pi(\mathbb{R}^2/\mathbb{Z}^2\times B) = Q(B)$ pour tout A, B sous-ensembles mesurables de $\mathbb{R}^2/\mathbb{Z}^2$.

Le problème de Kantorovich (1) peut être formulé sous sa forme duale, comme

$$\mathcal{T}_2(P,Q) = \sup_{(f,g)\in\Phi_2(P,Q)} \int f(\bar{x})dP(\bar{x}) + \int g(\bar{y})dQ(\bar{y}), \tag{2}$$

où

$$\Phi_2(P,Q) = \{(f,g) \in L_1(P)\times L_1(Q) :\ f(\bar{x}) + g(\bar{x}) \le d^2(\bar{x},\bar{x})\}.$$

On dit que $\psi \in L_1(P)$ est un *potentiel de transport optimal de P à Q pour le coût* d^2 s'il existe $\varphi \in L_1(Q)$ tel que le couple (ψ,φ) est une solution de (2).

Comme le tore plat est un espace compact, le carré du coût de transport $\mathcal{W}_2(P,Q) := \sqrt{\mathcal{T}_2(P,Q)}$ définit une distance qui métrise la convergence faible sur $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$. C'est pour cela que, quand on cherche à comparer des mesures sur le tore plat, l'étude des propriétés de ces mesures devient crucial. Concrètement, quand on considère l'équivalent empirique $\mathcal{T}_2(P_n,Q_n)$, on s'intéresse à sa vitesse de convergence. Les nombreux travaux récents sur cet objet ont montré que les bornes de l'espérance $E\mathcal{T}_2(P_n,Q_n)$ dépendent fortement de la dimension de l'espace (Fournier et al., 2015). Dans notre problème, la dimension est 2 et donc la vitesse de convergence et de l'ordre de $1/\sqrt{n}$, c'est-à-dire, il existe une constante C telle que $E\mathcal{T}_2(P_n,Q_n) \le C/\sqrt{n}$. Cependant, les constantes qui découlent de la démonstration de Fournier et al. (2015) sont trop grandes pour que ce résultat puisse être adapté en un test d'homogénéité. Une approche alternative consiste à tenir compte de l'inégalité de McDiarmid et de la formulation duale (2) pour montrer que

$$P(\mathcal{T}_2(P_n,Q_m) - E\mathcal{T}_2(P_n,Q_m) > t) \le \exp\left(-\frac{nm}{n+m}8t^2\right). \tag{3}$$

Finalement, deux résultats récents de del Barrio et Loubes (2019), et del Barrio, Gonzalez-Sanz et Loubes (2021) montrent que les fluctuations $\sqrt{n}(\mathcal{T}_2(P_n, Q_m) - E\mathcal{T}_2(P_n, Q_m))$ sont asymptotiquement gaussiennes pour des mesures sur $\mathbb{R}^d$ avec la distance euclidienne. On montre que ce résultat est aussi applicable dans l'espace $\mathbb{R}^2/\mathbb{Z}^2$, ce qui n'était pas couvert par les travaux précédents.

3 Tests d'homogénéité de deux échantillons pour des mesures sur $\mathbb{R}^2/\mathbb{Z}^2$

Les méthodes de test d'homogénéité basées sur la distance de Wasserstein restent encore un problème ouvert. Le cas d'un seul échantillon a été adressé par Hallin et al. (2021), mais des approches pour deux échantillons en dimension générale, et pour des mesures sur des espaces plus générales, n'ont pas encore été proposées. La difficulté intrinsèque de connaître la distribution de $\mathcal{W}_p(P_n, Q_m)$ explique l'absence de solutions, spécialement quand la dimension est supérieure à 1. Nous proposons plusieurs approches de test d'homogénéité basées sur la statistique $\mathcal{T}_p(P_n, Q_m) = \mathcal{W}_p^p(P_n, Q_m)$ et permettant l'évaluation de $H_0 : P = Q$ pour des mesures sur $\mathbb{R}^2/\mathbb{Z}^2$. Comme certaines de nos approches sont basées sur l'extension des resultats pour des mesures sur $\mathbb{R}^d$, elles devraient être aussi applicables à l'espace éuclidien de dimension générale. Si on note $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons i.i.d. de lois $P, Q \in \mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$ respectivement, et P_n, Q_m leur mesures de probabilité empiriques correspondantes, on cherche à tester l'hypothèse $H_0 : P = Q$ via la définition de la p-valeur

$$p = \mathrm{P}_{H_0}(\mathcal{T}_p(P_n, Q_m) \geq t_{nm}), \tag{4}$$

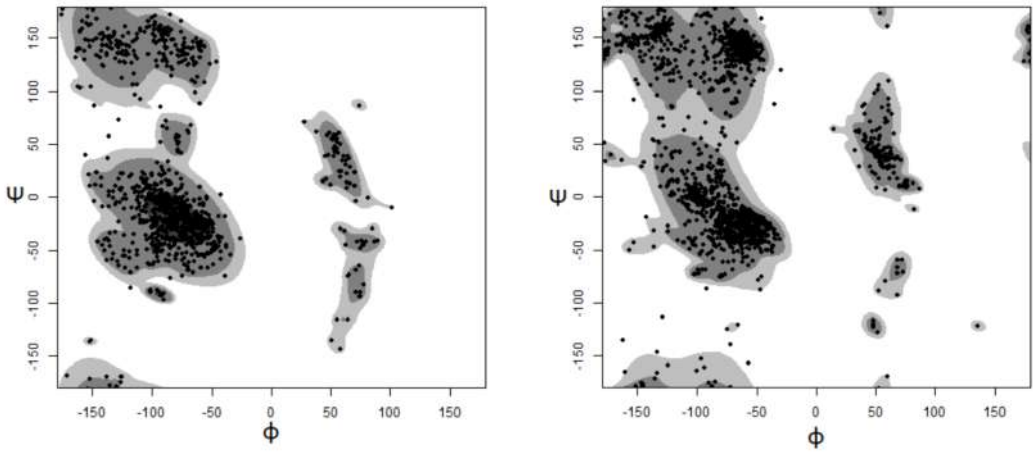
où t_{nm} désigne la réalisation de la statistique, i.e. la puissance d'ordre p de la distance de Wasserstein entre les mesures de probabilité empiriques correspondantes.
Comme on a remarqué, connaître la distribution de la statistique sous l'hypothèse nulle reste un problème ouvert (et peut être infaisable). Nous explorons donc des solutions approchées reposant sur:

1. la projection du problème sur un espace unidimensionnel,
2. la définition de bornes supérieures pour les p-valeurs (à partir des inégalités (3)),
3. le comportement asymptotique de la statistique sous l'hypothèse alternative.

La validité de tests obtenus par les trois approches envisagées a été évaluée empiriquement sur des données réelles de conformation de protéines, ainsi que sur des données simulées. Les résultats qui seront présentés montrent que les deux premières approches permettent de tester de manière efficace l'égalité de deux mesures sur $\mathbb{R}^2/\mathbb{Z}^2$, pour un

nombre de données suffisamment grand. La troisième approche ne permet pas la construction d'un test, ce qui s'explique car les lois du statistique sous les hypothèses nulle et alternative ne sont pas distinguables en pratique. Lors de la présentation, on montrera que les deux approches efficaces se complémentent entre elles, et comment on peut les combiner afin d'optimiser la performance computationnelle des analyses à implémenter.

Graphiques



ALA-GLN-CYS HIS-GLN-ARG

Figure 1: Exemples des distributions des angles (ϕ, ψ) correspondant aux acides aminés centraux dans les tripeptides (fragments de trois acides aminés) ALA-GLN-CYS (gauche) et HIS-GLN-ARG (droite).

Bibliographie

del Barrio, E. , Gonzalez-Sanz, A. and Loubes, J.-M. (2021). Central Limit Theorems for General Transportation Costs. *arXiv:2102.06379*.

del Barrio, E., Gordaliza, P. and Loubes, J.-M. (2019). A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning, *Information and Inference: A Journal of the IMA*, 8, 12.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.*, 47 (2) 926 - 951.

Fournier, N., and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields* 162, 707–738.

Ramdas, A., Garcia, N. and Cuturi, M. (2015). On wasserstein two sample testing and related families of nonparametric tests, *Entropy*, 19, 09.

Hallin, M., Mordant, G. and Segers, J. (2021). Multivariate goodness-of-fit tests based on wasserstein distance.

Generalisation bounds for deep neural networks

Benjamin Guedj

UCL Centre for Artificial Intelligence, 90 High Holborn, WC1V 6LJ London, United Kingdom
benjamin.guedj@inria.fr

Abstract. An increasing number of exciting results have recently contributed to bridge the gap between theoretical understanding of deep neural networks and their impressive empirical successes. I will focus on generalisation guarantees for (some) deep neural architectures, which in my view now lead to practical insights for machine learning practitioners.

Keywords. Generalisation, Deep Neural Networks.

Tests d'équivalence pharmacocinétique par modélisation : impact d'une mauvaise spécification du modèle

Mélanie Guhl [1], François Mercier [2], Satish Sharan [3], Kairui Feng [3], Guyoing Sun [4], Wanjie Sun [4], Stella Grosser [4], Liang Zhao [3], Lanyan Fang [3], France Mentré [1], Emmanuelle Comets [1,5] & Julie Bertrand [1]

[1] *Université de Paris, INSERM, IAME, UMR 1137, 75006 Paris, France - melanie.guhl@inserm.fr, france.mentre@inserm.fr, emmanuelle.comets@inserm.fr, julie.bertrand@inserm.fr*

[2] *Department of Biostatistics, Roche Innovation Center Basel, Basel, Switzerland - francois.mercier@roche.com*

[3] *Division of Quantitative Methods and Modeling, Office of Research Standards, Office of Generic Drugs,Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring MD 20993, USA - Satish.Sharan@fda.hhs.gov, Kairui.Feng@fda.hhs.gov, Liang.Zhao@fda.hhs.gov, Lanyan.Fang@fda.hhs.gov*

[4] *Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring MD 20993, USA - Guoying.Sun@fda.hhs.gov, Wanjie.Sun@fda.hhs.gov, Stella.Grosser@fda.hhs.gov*

[5] *INSERM, CIC 1414, Univ Rennes-1, 35700 Rennes, France*

Résumé. En cas de design épars, l'approche classique pour évaluer l'équivalence pharmacocinétique entre deux médicaments, basée sur une analyse non compartimentale, ne peut être mise en place. Dans ce cas, une alternative consiste à utiliser une approche par modélisation. Dans ce travail, nous nous intéressons à l'impact d'une mauvaise spécification du modèle pharmacocinétique. Au travers d'une étude de simulation inspirée d'une étude de biosimilarité entre deux formulations d'un anticorps monoclonal, nous montrons que la sélection du modèle est une étape clé dans les études d'équivalence pharmacocinétique par modélisation.

Mots-clés. Test d'équivalence, pharmacocinétique, modèles non linéaires à effets mixtes, design épars, biosimilarité

Abstract. In case of sparse design, the conventional approach to evaluate the pharmacokinetic equivalence between two drugs, based on a non-compartmental analysis, can be unfeasible. In that case, an alternative consists in using a model-based approach. In the present work, we investigate the impact of a misspecification of the pharmacokinetic model. Through a simulation study inspired by a biosimilarity study of two formulations of a monoclonal antibody, we show that model selection is a key step in model-based pharmacokinetic equivalence studies.

Keywords. Equivalence test, pharmacokinetics, non linear mixed effects models, sparse design, biosimilarity

1 Introduction

Lorsque l'on veut comparer les profils pharmacocinétiques (PK) de deux médicaments, la méthode classique est la comparaison des moyennes des aires sous la courbe (AUC) et des concentrations maximales (C_{max}), grâce à un test statistique appelé le Two One Sided Tests (TOST). Classiquement, ces paramètres PK (AUC,C_{max}), sont obtenus grâce à une approche non compartimentale (NCA), comme recommandé par la FDA (2013).

Lorsque l'AUC et la C_{max} ne peuvent être calculés par NCA en raison du design épars de l'étude, une alternative proposée par Dubois et. al. (2011) est de mettre en œuvre une approche par modélisation (MB). L'approche MB consiste à inférer sur les paramètres d'un modèle non linéaire PK à effets mixtes.

Dans ce travail, nous avons évalué, par simulations, l'impact du modèle PK et des effets traitement estimés sur les résultats de l'approche MB. Nos simulations s'inspirent d'une étude de biosimilarité entre deux formulations d'un anticorps monoclonal développé par le laboratoire Roche. En raison de la longue demi-vie du médicament, l'essai clinique a été mené avec un design parallèle.

2 Méthodes

2.1 Le Two One Sided Test

Le Two One Sided Tests (TOST) consiste à mettre en œuvre deux tests de Wald comparant l'effet traitement β^T à une valeur seuil δ. La Food and Drug Administration (FDA) fixe ce seuil à $\delta = log(1.25)$. L'hypothèse nulle de non-équivalence est :

$$H_0 : \{\beta^T \leq -\delta \text{ or } \beta^T \geq \delta\} \tag{1}$$

L'effet traitement sur AUC et C_{max} est défini comme l'espérance des différences en log des AUC et C_{max} individuelles dans le bras référence et le bras test :

$$\beta^T_{AUC} = \mathbb{E}(log(\frac{AUC_T}{AUC_R})) \tag{2}$$

$$\beta^T_{Cmax} = \mathbb{E}(log(\frac{Cmax_T}{Cmax_R})) \tag{3}$$

Les résultats sont souvent présentés par le biais du ratio des moyennes géométriques (GMR) qui est l'exponentielle de β^T et dont l'intervalle à 90% doit être compris dans l'intervalle [0.8; 1.25] pour que l'hypothèse nulle soit rejetée.

2.2 Étude réelle

Les données utilisées proviennent de deux essais cliniques randomisés mis en place dans le cadre de l'investigation de la biosimilarité entre deux formulations d'un anticorps monoclonal. La première étude (S1) est composée de cinq bras parallèles de 24 patients : trois bras référence à différents niveaux de dose (105, 225 et 300mg) et deux bras test (105 et 225mg). Dans la seconde étude (S2), composée d'un bras référence de 25 patients et d'un bras test de 23 patients, la dose testée est 225mg. Les patients ont été suivis pendant 13 semaines.

Les analyses ont été faites séparément pour chaque étude et dose testée. Sur les données complètes (11 points par sujet), pour l'approche MB, le modèle PK structurel et le modèle d'erreur résiduelle ont été sélectionnés sur la base du BIC. Nous avons comparé les résultats des tests obtenus par l'approche MB avec les résultats obtenus par NCA (NCA-TOST). Pour l'approche MB, l'erreur standard de β^T est dérivée de la matrice d'information de Fisher, borne inférieure de l'estimateur de la matrice de variance-covariance d'estimation à l'asymptotique (MB-TOST Asympt).

Les analyses ont ensuite été menées sur un sous-ensemble plus épars des données. Ces protocoles épars (5 points par sujet) ont été obtenus par optimisation du design avec $PFIM$, à partir des modèles obtenus sur les données complètes. Dans ce cas de figure, la méthode NCA n'est pas applicable, et les erreurs standard asymptotiques peuvent ne pas refléter correctement la variabilité présente dans les données. Nous avons donc testé l'équivalence PK uniquement par modélisation, en comparant trois méthodes différentes de calcul des erreurs standards : la méthode asymptotique (MB-TOST Asympt), la correction de Gallant proposée par Bertrand et. al. (2012) (MB-TOST Gallant) et la génération de distribution a posteriori à partir d'un algorithme HMC, proposée par Loingeville et. al. (2020) (MB-TOST Post).

2.3 Étude de simulation

Ce cas réel a inspiré une étude de simulations, avec design riche et épars. Sur les simulations riches, nous avons comparé les performances de NCA-TOST et MB-TOST Asympt en termes d'erreur de type I et de puissance, en explorant l'impact de la modélisation des effets traitement. Sur les simulations éparses, nous avons comparé les performances de MB-TOST Asympt, MB-TOST Gallant et MB-TOST Post en termes d'erreur de type I et de puissance, en explorant l'impact du modèle PK (1 versus 2 compartiments de distribution) utilisé, et la pertinence d'une étape de sélection du modèle PK sur le bras référence en amont de l'étude d'équivalence.

3 Résultats

3.1 Étude réelle

Les résultats des approches NCA et MB sont concordantes. Sur les données riches, le modèle PK le plus adapté est un modèle à deux compartiments avec absorption et élimination linéaires. Sur les données éparses, le modèle PK choisi est un modèle à un seul compartiment. Sur les données riches comme éparses, la biosimilarité des deux formulations n'a pas pu être montrée (Figure 1).

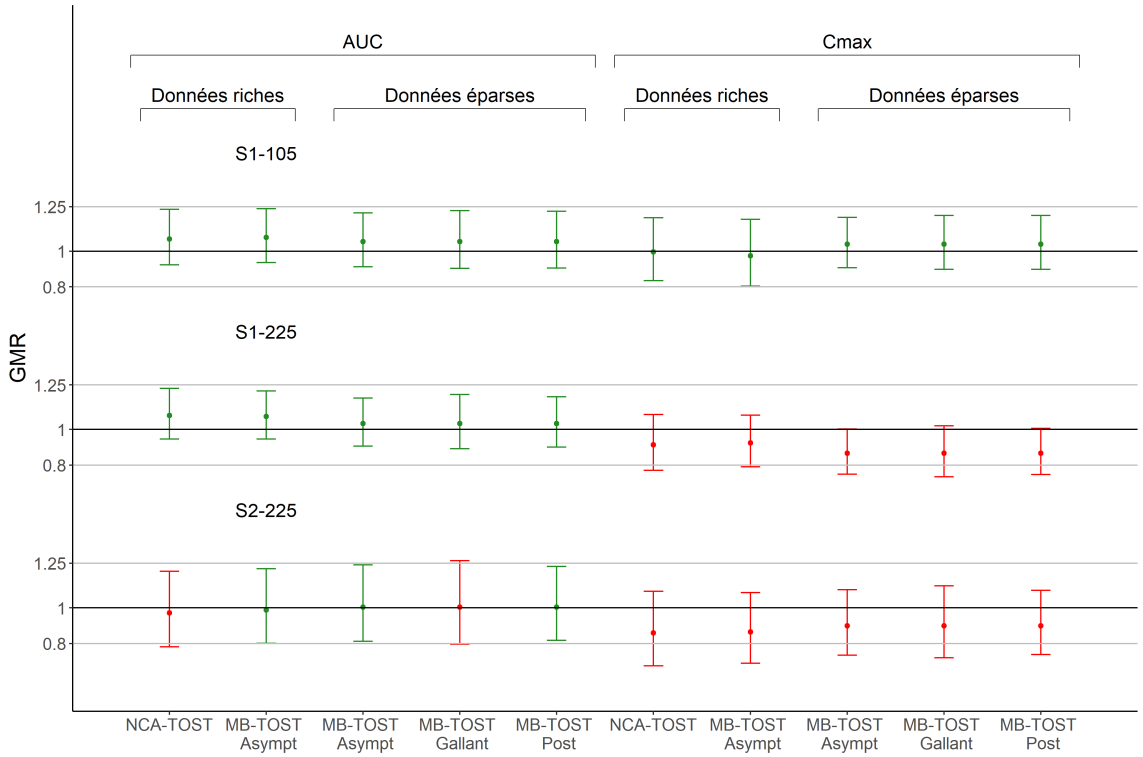


Figure 1: Ratios des moyennes géométriques (GMR) et leurs intervalles de confiance à 90% pour AUC et C_{max}, avec NCA-TOST et MB-TOST Asympt sur les données riches et avec MB-TOST Asympt, Gallant et Post sur les données éparses. Les lignes grises sont les limites de l'intervalle $\mathcal{H}_1$, $GMR = 0.8$ et $GMR = 1.25$, et la ligne noire représente $GMR = 1$.

3.2 Étude de simulation

Le modèle PK qui a été utilisé pour simuler les données est celui sélectionné pour décrire les données réelles du bras référence de S1 à la dose 225mg, correspondant à un modèle à deux compartiments avec absorption et élimination linéaires, avec des effets traitement simulés sur la clairance (Cl) et le volume (V_1) du compartiment principal.

Sur les simulations avec un protocole riche, lorsque l'on travaille avec le modèle PK structurel simulé en estimant des effets traitement sur tous les paramètres apparents, les erreurs de type I obtenues avec MB-TOST Asympt sont similaires à celles obtenues avec NCA-TOST et proches de la valeur nominale de 5%. Une modélisation des effets traitement différente de celle utilisée pour simuler les données (effets traitement estimés sur la constante d'absorption ka et la biodisponibilité F) ne donne pas de résultats satisfaisants.

Sur les simulations avec un protocole épars, MB-TOST Asympt et Post permettent de contrôler les erreurs de type I avec le modèle PK utilisé pour simuler les données. Dans notre exemple où l'éloignement à l'asymptotique est faible, MB-TOST Gallant se révèle trop conservateur. Lorsque le modèle PK est mal spécifié (un seul compartiment estimé), on observe une inflation de l'erreur de type I sur C_{max}.

Enfin sur les simulations éparses, si on ajoute une étape de sélection sur la base du BIC du nombre de compartiments du modèle PK sur les données de référence, le bon modèle est sélectionné dans la majorité des cas et l'erreur de type I des tests est contrôlée (Figure 2).

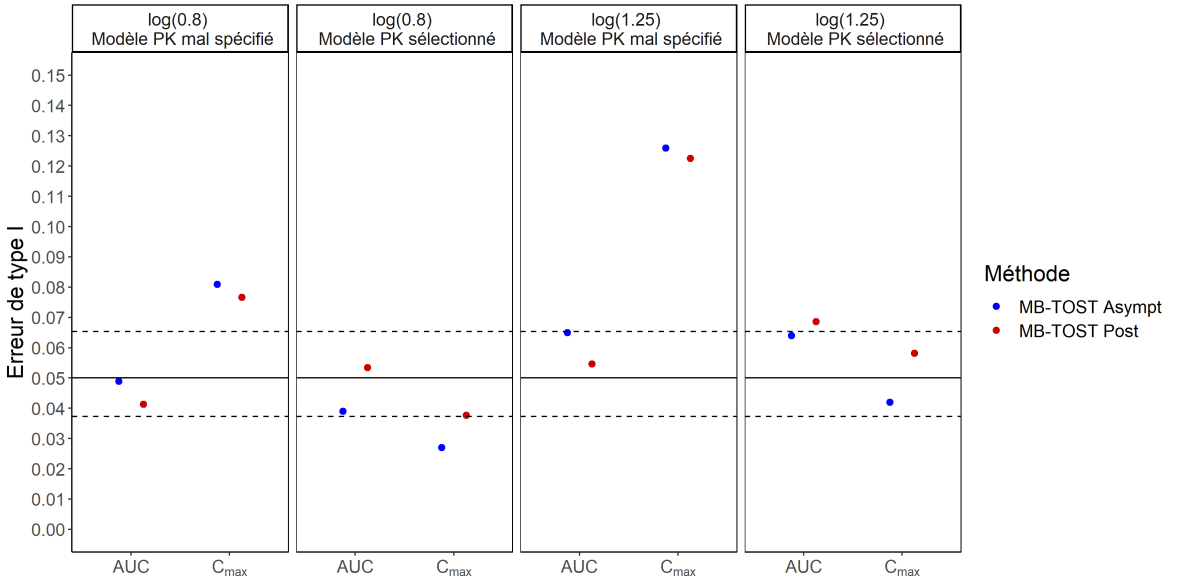


Figure 2: Erreurs de type I pour MB-TOST Asympt et MB-TOST Post sur les données éparses simulées avec effets traitements simulés sur AUC et C_{max} égaux à $log(0.8)$ et $log(1.25)$. Les lignes en pointillés représentent l'intervalle de confiance à 95% autour de 0.05 pour 1000 jeux de données.

4 Conclusion

L'approche par modélisation apparaît comme une alternative robuste à l'approche non compartimentale dans le cas de designs épars. Notre étude de simulation a permis de montrer que la sélection du modèle PK est une étape clé dans la mise en œuvre d'une approche par modélisation pour les études d'équivalence PK. Cette approche, utilisée ici pour montrer la biosimilarité, peut aussi être appliquée aux études de bioéquivalence où l'on compare un générique à un médicament de référence.

Bibliographie

U.S. Food and Drug Administration (2013). Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA. `https://www.fda.gov/media/87219/download`

A Dubois, M Lavielle, S Gsteiger, E Pigeolet et F Mentré (2011). Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm, *Statistics in Medicine*, 30(21):2582–2600

J Bertrand, E Comets, M Chenel et F Mentré (2012). Some alternatives to asymptotic

tests for the analysis of pharmacogenetic data using nonlinear mixed effects models, *Biometrics*, 68(1):146–155, 201

F Loingeville, J Bertrand, TT Nguyen, S Sharan, G Sun, S Grosser, L Zhao, L Fang, K Möllenhoff, H Dette et F Mentré (2020). New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling, *The AAPS Journal, in press*.

COVAL NANCY - ETUDE DE SÉROPRÉVALENCE CONTRE LE VIRUS SARSCOV-2 (COVID-19) DANS LA POPULATION DE LA MÉTROPOLE DU GRAND NANCY

Anne GÉGOUT PETIT1[1]; Hélène JEULIN[1,2]; Karine LEGRAND[1]; Agathe BOCHNAKIAN[4]; Pierre VALLOIS[1]; Evelyne SCHVOERER[2,3]; Francis GUILLEMIN[4]

[1]*Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France, anne.gegout-petit@univ-lorraine.fr ; pierre.vallois@univ-lorraine.fr*
[2]*Université de Lorraine, CNRS, LCPME, F- 54000 Nancy, France, e.schvoerer@chru-nancy.fr ; h.jeulin@chru-nancy.fr*
[3]*Laboratoire de Virologie, CHRU de Nancy Brabois, F- 54500 Vandoeuvre- lès- Nancy, France,*
[4]*CHRU -Nancy, INSERM, Université de Lorraine, CIC Epidémiologie clinique, F-54000 Nancy, France, k.legrand@chru-nancy.fr ; A.BOCHNAKIAN@chru-nancy.fr ; francis.guillemin@chru-nancy.fr*

Résumé. Dès le début de la pandémie mars 2020, l'Organisation Mondiale de la Santé a recommandé des études de séroprévalence sur un échantillon aléatoire de la population afin de préciser les connaissances sur l'étendue de la circulation du SARS-CoV-2 et d'évaluer l'immunité acquise d'une population ; cette recommandation a d'ailleurs été relayée par le président de la SFdS, Jean-Michel Marin et la plateforme MODCOVID de l'INSMI. Nous présenterons l'histoire de ce projet à l'initiative des mathématiciens mais qui n'aurait pas pu se réaliser sans un financeur et sans l'implication d'épidémiologistes et/ou virologues du CHRU de Nancy. Bien sûr nous donnerons quelques éléments de la méthode et les résultats.

Mots-clés. COVID-19, séroprevalence, échantillon aléatoire, facteurs de risque, profil de symptômes, transmission intra foyer, séroneutralisation

Abstract. From the start of the pandemic in March 2020, the World Health Organisation recommended seroprevalence studies on a random sample of the population in order to clarify knowledge on the extent of circulation of SARS-CoV-2 and to evaluate the acquired immunity of a population; this recommendation was also relayed by the president of the SFdS, Jean-Michel Marin and the INSMI's MODCOVID platform. We will present the history of the project, which was initiated by mathematicians but could not have been carried out without a funder and the involvement of epidemiologists and/or virologists from the Nancy CHRU. Of course we will give some elements of the method and the results.

Keywords. COVID-19, seroprevalence, random sample, risk factors, symptoms profiles, seroneutralisation

1. Méthode

6094 individus identifiés par tirage au sort sur les listes électorales, stratifié sur les zones IRIS de la MGN ont été invités avec tous les membres de leur foyer, âgés de plus de 5 ans, à une visite réalisée entre le 26 juin et le 24 juillet 2020. Cette visite inclut le remplissage d'un questionnaire explorant les caractéristiques sociodémographiques et le niveau de précarité sociale, médicales, les contacts potentiels avec le virus de la COVID-19, le recensement de 18 symptômes avec leur intensité et la réalisation d'un prélèvement sanguin. Un test ELISA (Bio-rad) a été utilisé pour détecter les anticorps anti-SARS-CoV-2 (IgT, c'est-à-dire IgA/IgG/IgM). Les échantillons de sérum ont été classés en fonction de l'activité de séroneutralisation >50% (NT50). Chaque zone IRIS a été associée au score EDI (indice écologique de défavorisation sociale).

Les analyses statistiques étaient relativement standard (calcul de pourcentages, ajustement, redressement pour la transposition de la prévalence à l'échelle de la métropole ou de la France ; régression logistique et Odds Ratio, test du chi-deux ou de Fisher pour la comparaison des groupes). Nous avons aussi utilisé des méthodes de clustering de variables pour la description des profils de symptômes ainsi qu'un test et des méthodes de simulation par permutations pour tester la transmission et estimer le risque relatif d'être infecté lorsque l'on est dans un foyer avec un membre séropositif.

2. Résultats

Parmi les 2006 participants âgés de 5 à 95 ans dont 55% de femmes et 148 mineurs, 21% sont considérés en « précarité sociale » (score Epices >30). Parmi les participants, 252 (12,6%) pensaient avoir été infectés par COVID-19 parce qu'ils avaient ressenti des symptômes (86%) et/ou avaient été en contact avec une personne malade (44%).

43 participants étaient séropositifs au SARS-CoV-2 soit une séroprévalence brute de 2,1% (IC 95% [1,5-2,9]) et à 2,30% après standardisation selon l'âge et le sexe en France. Elle était plus élevée pour les 20-34 ans (4,7 %, IC95% [2,3 - 8,4]), dans les EDI associé à niveau socio-économique inférieur (2,7% pour quintiles 3-4-5 contre 1 % pour les quintiles 1-2, p=0,02). Cependant, elle n'est pas significativement plus faible chez les personnes en situation de précarité sociale (1,0% pour les scores EPICES >30 contre 2,5% sinon, p=0,09). La transmission intra-foyer était significative ($p=10^{-6}$) avec un RR = 30 (IC95% =[20 ; 78]).

En ce qui concerne les symptômes 25% des individus (IC95% [23 - 27]) ont déclaré au moins un des 4 symptômes de la COVID-19[1], ce critère est significativement lié à la séroprévalence (6,5% contre 0,7%, $p<10^{-13}$). Près de la moitié des individus (47% (resp. 14%)) ont déclaré avoir ressenti au moins un des symptômes (resp. un symptôme "intense"). La présence d'au moins un symptôme était significativement lié à une séroprévalence plus élevée (3,8 % contre 0,7 %, $p<10^{-5}$). Cette différence est accentuée lorsqu'au moins un des symptômes est qualifié d'intense (9,4 % contre 0,7 %, $p<10^{-17}$). L'anosmie ou l'agueusie (perte d'odorat ou de gout) est le symptôme le plus discriminant (OR=27,8, CI= [13,9 - 54,5]).

16,3 % des personnes séropositives étaient totalement asymptomatiques alors que les

[1]D'après la définition de l'ECDC

symptomatiques ont déclaré des symptômes dans les trois première dernières semaines de mars ; montrant un net effet du confinement dans la diminution de la propagation de la maladie. Enfin, pour 31/43 (72%, [56 - 85]) séropositifs, la détection des anticorps a été associée à une activité de neutralisation du SARS-CoV-2 démontrée in vitro au laboratoire.

3. Conclusion

Cette étude met en évidence une très faible prévalence des sérologies positives anti-SARS-CoV-2, lors de la 1ère vague laissant supposer un effet bénéfique du confinement, avec une séroneutralisation fréquente du SARS-CoV-2 chez les patients IgT-positifs. Des études supplémentaires réalisées sur des populations variées devront préciser la durée de persistance de cette neutralisation sérique, sa corrélation avec la protection des individus contre l'infection et/ou la sévérité de la maladie COVID-19, et une éventuelle protection croisée contre des souches de SARS-CoV-2 qui évoluent sur le plan génétique.

Bibliographie

Organization WH (2020). Population-based age-stratified seroepidemiological investigation protocol for coronavirus 2019 (COVID-19) infection.
https://apps.who.int/iris/handle/10665/332188 (May, 2020 date last accessed)

Case definition for coronavirus disease 2019 (COVID-19), as of 3 December 2020. European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/en/covid-19/surveillance/case-definition (Jan, 2021 date last accessed)

Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. (2012) ClustOfVar: An R Package for the Clustering of Variables. *J Stat Soft*. 50(13). Disponible sur: http://www.jstatsoft.org/v50/i13/

Siegel S, Castellan NJ. (1988) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill; 440 p.

Apprentissage de modèles CHARME avec des réseaux de neurones

José G. Gómez-García [1] · Jalal Fadili [2] · Christophe Chesneau [3]

[1] *Université Paris-Saclay, AgroParisTech, UMR 518 MIA-Paris, INRAE. AgroParisTech, 16 rue Claude Bernard, 75231 Paris.*
jose.gomez-garcia@agroparistech.fr

[2] *Normandie Université, ENSICAEN, UNICAEN, GREYC, UMR CNRS 6072. ENSICAEN, 6 Bd du Maréchal Juin, 14050 Caen.*
Jalal.Fadili@ensicaen.fr

[3] *Normandie Université, UNICAEN, LMNO, UMR CNRS 6139. LMNO, Sciences 3, Campus 2, Bd du Maréchal Juin, 14000 Caen.*
christophe.chesneau@unicaen.fr

Résumé. Dans cette note, nous considérons un modèle appelé CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). En quelques mots, c'est un modèle de mélange généralisé de séries chronologiques non linéaire et non paramétrique AR-ARCH. Nous garantissons la stabilité (ergodicité et stationnarité) du modèle sous certaines conditions de type Lipschitz pour les fonctions d'autorégression et de volatilité, lesquelles sont beaucoup plus faibles que celles présentées dans la littérature existante. Ce résultat et la propriété d'approximation universelle de réseaux de neurones (RN), possiblement avec des architectures profondes (RNP), nous fournit les bases pour développer une théorie d'apprentissage pour les fonctions d'autorégression-basées-sur-RN du modèle. En outre, la consistance forte et la normalité asymptotique de l'estimateur des poids et des biais des RN considéré sont garanties sous de faibles conditions.

Mots-clés. modèle AR-ARCH non-paramétrique ; réseaux de neurones profonds ; modèles de mélange ; séquence à changement de régime markoviens ; dépendance τ-faible ; ergodicité ; stationnarité ; identifiabilité ; consistance ; signaux d'EEG.

Abstract. In this note, we consider a model called CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). Roughly speaking, this is a class of generalized mixture of nonlinear nonparametric AR-ARCH time series. We guarantee the stability (ergodicity and stationarity) of the model under certain Lipschitz-type conditions on the autoregression and volatility functions, which are much weaker than those presented in the current literature. This result and the universal approximation property of neural networks (NN), possibly with deep architectures (DNN), provides us with the bases for developing a learning theory for the NN-based autoregressive functions of the model. By the way, the strong consistency and asymptotic normality of the considered estimator of the NN weights and biases are guaranteed under weak conditions.

Keywords. Nonparametric AR-ARCH ; deep neural network ; mixture models ; Markov switching ; τ-weak dependence ; ergodicity ; stationarity ; consistency ; EEG signals.

1 Introduction

Dans l'analyse de séries chronologiques, il est commun d'étudier les modèles tels que : AR, ARMA, ARCH, GARCH, etc. ; ou plus généralement, le modèle CHARN

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, \theta^0) + g(X_{t-1}, \ldots, X_{t-p}, \lambda^0)\epsilon_t, \qquad t \in \mathbb{Z}, \tag{1}$$

où f, g sont des fonctions inconnues et $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc indépendant. Cependant, dans la pratique, il n'est pas toujours réaliste de supposer que le processus observé ait la même tendance f et la même volatilité g à chaque instant t. Entre autre, c'est le cas des signaux d'EEG, voir Lo *et al.* (2009), où l'on peut observer des changements de comportement, même brusques, lesquelles on ne peut pas les modéliser même en utilisant les modèles localement stationnaires. C'est pour cela que nous nous concentrons sur un modèle plus général, appelé CHARME, qui prend en compte ces changement brusques de comportement.

Pour définir ce modèle, considérons l'espace de Banach $(E, \|\cdot\|)$, doté de sa tribu borélienne $\mathcal{E}$. L'espace produit E^p est alors naturellement doté de sa tribu produit $\mathcal{E}^{\otimes p}$. Le modèle **CHARME(p)**, à valeurs dans E, est la série chronologique définie par

$$X_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k^0) + g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k^0)\epsilon_t \right) \quad t \in \mathbb{Z}, \tag{2}$$

où

- pour chaque $k \in [K] := \{1, 2, \ldots, K\}$, $f_k : E^p \times \Theta_k \longrightarrow E$ et $g_k : E^p \times \Lambda_k \longrightarrow \mathbb{R}$ sont respectivement les fonctions d'autorégression et de volatilité, avec des espaces de paramètres respectifs Θ_k et Λ_k, $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Theta_k)$- et $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Lambda_k)$- mesurables, où $\mathcal{B}(\Theta_k)$ est la tribu borélienne sur Θ_k et pareillement pour Λ_k ;
- $(\epsilon_t)_t$ est un bruit blanc indépendant à valeurs dans E ;
- $\xi_t^{(k)} = \mathbb{I}_{\{R_t=k\}}$, avec $\mathbb{I}_{\mathcal{C}}$ désignant la fonction caractéristique de $\mathcal{C}$ (*i.e.*, elle vaut 1 sur $\mathcal{C}$ et 0 sinon), où $(R_t)_{t\in\mathbb{Z}}$ est une séquence de variables aléatoires indépendantes à valeurs dans l'espace fini $[K]$, qui est en plus indépendante du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. Par la suite, on pose $\pi_k = \mathbb{P}(R_0 = k)$.

Le modèle (2) peut être étendu au cas $p = \infty$. Nous l'appellerons alors modèle CHARME à mémoire infinie et que nous désignerons par CHARME(∞). Dans ce cadre, l'espace d'états du modèle est le sous-ensemble de $E^{\mathbb{N}}$:

$$E^{\infty} := \left\{ (x_k)_{k>0} \in E^{\mathbb{N}} : \ x_k = 0 \text{ for } k > N, \text{ for some } N \in \mathbb{N}^* \right\},$$

doté de sa tribu produit $\mathcal{E}^{\otimes\mathbb{N}}$.

Il est clair que le modèle (2) contient le modèle (1) (cela correspond au cas $K = 1$ en (2)). D'ailleurs, des applications de ce modèle ont été traitées d'une manière directe ou indirecte dans plusieurs domaines de recherche. Voir par exemple : Tadjuidje-Kamgaing, J. (2005), Weigend, A.S. and Shi, S. (2000), Kirch, C. and Kamgaing, T. (2012) et Liehr *et al.* (1999).

2 Ergodicité et stationnarité des modèles CHARME

Le résultat suivant nous fournit des conditions pour avoir la stabilité du modèle dont la preuve est donnée dans la Section 8.1 de Gómez-García *et al.* (2020).

Théorème 1. *Considérons le modèle CHARME(∞), i.e.,* (2) *avec* $p = \infty$. *Supposons qu'il existe des séquences non-négatives* $(a_i^{(k)})_{i\geq 1,k\in[K]}$ *et* $(b_i^{(k)})_{i\geq 1,k\in[K]}$ *telles que, pour tout* $x, y \in E^\infty$ *et tout* $k \in [K]$,

$$\|f_k(x,\theta_k^0) - f_k(y,\theta_k^0)\| \leq \sum_{i=1}^{\infty} a_i^{(k)}\|x_i - y_i\|, \quad |g_k(x,\theta_k^0) - g_k(y,\theta_k^0)| \leq \sum_{i=1}^{\infty} b_i^{(k)}\|x_i - y_i\| \quad (3)$$

Notons $A_k = \sum_{i=1}^{\infty} a_i^{(k)}$, $B_k = \sum_{i=1}^{\infty} b_i^{(k)}$ *et* $C(m) = 2^{m-1}\sum_{k=1}^{K}\pi_k\left(A_k^m + B_k^m\|\epsilon_0\|_m^m\right)$. *Alors, nous obtenons les affirmations suivantes :*

(i) si $c := C(1) < 1$, *alors il existe une solution strictement stationnaire* $(X_t)_{t\in\mathbb{Z}}$ *du modèle CHARME(∞) appartenant à* $\mathbb{L}^1$.

(ii) si en plus $C(m) < 1$ *pour certain* $m > 1$, *alors cette solution appartient à* $\mathbb{L}^m$.

Remarque 1.

(1.1) Le résultat précédent est également valable dans le cas $p < \infty$. En effet, il suffit de prendre $a_i^{(k)} = b_i^{(k)} = 0$ pour tout $i > p$ et tout $k \in [K]$ dans les inégalités (3).

(1.2) Remarquons que le modèle CHARME(∞) (2) avec $p = \infty$ peut être réécrit comme une séquence de Markov $X_t = F(X_{t-1}, X_{t-2}, \ldots; \tilde{\xi}_t)$, $t \in \mathbb{Z}$, via la fonction

$$F(x; (\xi^{(0)}, \ldots, \xi^{(K)})) = \sum_{k=1}^{K} \xi^{(k)}\left(f_k(x,\theta_k^0) + g_k(x,\lambda_k^0)\xi^{(0)}\right), \quad (4)$$

avec des innovations $\tilde{\xi}_t := (\epsilon_t, \xi_t^{(1)}, \ldots, \xi_t^{(K)}) = (\epsilon_t, \xi_t) \in E\times B_e$, où $B_e := \{e_1, \ldots, e_K\}$ est la base canonique de $\mathbb{R}^K$. Sous les hypothèses du Théorème 1, la fonction F est continue car les fonctions $f_k(\cdot,\theta_k^0)$ et $g_k(\cdot,\lambda_k^0)$ sont continues par la condition (3). Il découle alors de (Doukhan, P. and Wintenberger, O, 2008, Lemma 5.5) et de la complétude de $\mathbb{L}^m$, qu'il existe une fonction mesurable H telle que le processus CHARME(∞) peut être écrit comme $X_t = H(\tilde{\xi}_t, \tilde{\xi}_{t-1}, \ldots)$. C'est-à-dire : le processus CHARME(∞) peut être représenté par un décalage de Bernoulli causal. En outre, sous ces hypothèses, $(X_t)_{t\in\mathbb{Z}}$ est le seul décalage de Bernouli causal, solution à (2) avec $p = \infty$. Donc, la solution $(X_t)_{t\in\mathbb{Z}}$ est automatiquement un processus ergodique. Enfin, le théorème ergodique implique la LFGN pour ce processus. Cette conséquence du Théorème 1 sera un résultat clé pour établir la consistance forte lorsqu'il s'agit d'estimer les fonctions d'autorégression et de volatilité du modèle CHARME(p).

(1.3) Stockis *et al.* (2010) montre l'ergodicité du modèle CHARME(p) avec $p < \infty$, sous réserve de multiple conditions. En particulier, les auteurs demandent la régularité du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. En revanche, nous n'avons pas besoin de cette restriction ici.

3 Estimation des paramètres du modèle : consistance

Soit $(X_t)_{1-p\leq t\leq n}$ $n+p$ observations de la solution strictement stationnaire $(X_t)_{t\in\mathbb{Z}}$ du modèle (2) (cette solution existe grâce au Théorème 1). Supposons que le nombre d'états K est connu et que nous avons accès aux observations des variables cachées iid $(R_t)_{1-p\leq t\leq n}$, ou bien, des variables $(\xi_t^{(k)})_{1-p\leq t\leq n, k\in[K]}$. Une hypothèse similaire peut être trouvée dans la littérature pour des cas spécieux du modèle CHARME. Voir, *e.g.*, Tadjuidje-Kamgaing, J. (2005) et Stockis *et al.* (2010).

Notre objectif est d'étudier un estimateur non linéaire des paramètres

$$(\theta^0,\lambda^0) := (\theta_1^0,\ldots,\theta_K^0,\lambda_1^0,\ldots,\lambda_K^0)$$

du modèle CHARME(p) (2) à partir des observations $(X_t)_{1-p\leq t\leq n}$ et $(\xi_t^{(k)})_{1-p\leq t\leq n,k\in[K]}$. Cet objectif est atteint en résolvant le problème de minimisation

$$\begin{gathered}(\widehat{\theta}_n,\widehat{\lambda}_n)\in \operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} Q_n(\theta,\lambda), \text{ où}\\ Q_n(\theta,\lambda) := \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\xi_t^{(k)}\ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big).\end{gathered} \tag{5}$$

Ici, $\ell : E\times E\times\mathbb{R}\longrightarrow\mathbb{R}\cup\{+\infty\}$ est une certaine fonction de coût. En général, ℓ devrait satisfaire $\ell(u,u,\tau)=0$, $\forall\tau$.

Afin de présenter notre résultat de consistance, il est plus commode de définir les processus

$$Y_t=(X_{t-p},X_{t-p+1},\ldots,X_t) \quad \text{et} \quad \xi_t=(\xi_t^1,\ldots,\xi_t^K), \quad t\in\mathbb{Z}.$$

Soit $(E^{p+1}\times B_e,\mathcal{E}^{\otimes(p+1)}\otimes\Xi,P)$ l'espace de probabilité commun, dans lequel sont définis les vecteurs aléatoires Y_t et ξ_t. Adoptons la notation suivante :

$$h(Y_t,\xi_t,\theta,\lambda) := \sum_{k=1}^{K}\xi_t^{(k)}\ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big). \tag{6}$$

En utilisant des arguments complexes du calcul des variations (en particulier sur les intégrands normaux et l'épi-convergence (voir Rockafellar, R.T. (1976) et Rockafellar, R.T. and Wets, R.J.B. (1998)), nous pouvons établir la consistance de l'estimateur (5) sous de faibles conditions. En particulier, sans la nécessité d'avoir un échantillon iid ni la différentiabilité de la fonction Q_n. Ceci est résumé dans le théorème suivant :

Théorème 2. *Soit $(X_t)_{t\in\mathbb{Z}}$ une solution strictement stationnaire et ergodique du modèle* (2) *(elle existe grâce au Théorème 1 avec $C(m)<1$ pour certain $m\geq 1$). Considérons les conditions raisonnables (***A.1***)-(***A.7***) de Gómez-García et al. (2020). Alors,*

(i) chaque point d'accumulation de $(\widehat{\theta}_n,\widehat{\lambda}_n)_{n\in\mathbb{N}}$ appartient à $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda}\mathbb{E}h(Y,\xi,\theta,\lambda)$ p.s.

(ii) si de plus la suite $(Q_n)_{n\in\mathbb{N}}$ est équi-coercitive, et $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda}\mathbb{E}h(Y,\xi,\theta,\lambda)=\{\theta^0,\lambda^0\}$, alors $(\widehat{\theta}_n,\widehat{\lambda}_n)\to(\theta^0,\lambda^0)$ et $Q_n(\widehat{\theta}_n,\widehat{\lambda}_n)\to\mathbb{E}h(Y,\xi,\theta^0,\lambda^0)$ p.s.

4 Apprentissage du modèle avec des RNP

Il est connu que, étant donné toute fonction cible continue f et une précision cible $\epsilon > 0$, les réseaux de neurones (RN) avec suffisamment paramètres (poids et biaises) judicieusement choisis donnent une approximation de la fonction pour une erreur de taille ϵ. Cette propriété d'approximation universelle des RN, nous permet de considérer le modèle CHARME(p) (2) avec les fonctions f_k et g_k exactement modélisées par des RN, avec $E = \mathbb{R}^d$. Pour une introduction de réseaux de neurones (profonds), voir Section 2.2 de Gómez-García *et al.* (2020). Avec les mêmes notations de cette dernière section, pour chaque $k \in [K]$, soit $\theta_k = \Big((W_k^{(1)}, b_k^{(1)}), \ldots, (W_k^{(L_k)}, b_k^{(L_k)})\Big)$, où $W_k^{(l)}$ et $b_k^{(l)}$ sont respectivement la matrice des poids et le vecteur de biaises de la l-ème couche du RN f_k. Similairement $\lambda_k = \Big((\bar{W}_k^{(1)}, \bar{b}_k^{(1)}), \ldots, (\bar{W}_k^{(\bar{L}_k)}, \bar{b}_k^{(\bar{L}_k)})\Big)$ pour le RN g_k. De plus, nous considérons la même fonction d'activation φ pour toutes les couches des RN f_k, g_k, avec $k \in [K]$.

Ergodicité et Stationnarité. En considérant les notations du Théorème 1, les précédentes notations et en notant $\|W_k^l\|$ la norme spectrale de la matrice correspondante, on peut démontrer que

$$A_k = \big(\mathrm{Lip}(\varphi)^{L_k - 1} \prod_{l=2}^{L_k} \|W_k^{(l)}\|\big) \sum_{i=1}^{p} \|W_{k,i}^{(1)}\| \quad \text{et} \quad B_k = \big(\mathrm{Lip}(\varphi)^{\bar{L}_k - 1} \prod_{l=2}^{\bar{L}_k} \|\bar{W}_k^{(l)}\|\big) \sum_{i=1}^{p} \|\bar{W}_{k,i}^{(1)}\|.$$

Par conséquence, si $C(m) = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m\right) < 1$ pour un certain $m \geq 1$, il existe une solution strictement stationnaire du modèle CHARME(p)-basé-sur-RN.

Consistance. Les conditions (**A.1**)-(**A.7**) de Gómez-García *et al.* (2020) sont satisfaites pour les RN f_k et g_k, pour tout $k \in [K]$ (cela a été montré dans le cité article). Donc, l'existence de la solution stationnaire et ergodique du modèle CHARME(p)-basé-sur-RN, implique le Théorème 2(i).

Pour pouvoir appliquer le Théorème 2(ii), nous avons besoin d'une certaine équi-coercitivité et unicité des vrais paramètres (θ^0, λ^0). Ceux-ci sont discutées et assurées dans la Section 6.2.1 de Gómez-García *et al.* (2020).

5 Commentaires

Établir la normalité asymptotique de l'estimateur (5) est très complexe dans un cadre variationnel. C'est pour cela que nous nous restreignons aux arguments habituels de la théorie d'inférence statistique qui demandent, en particulier, la dérivabilité d'ordre trois de la fonction Q_n. En plus, pour simplifier les résultats, nous prenons $\ell(u, v, \tau) = \|u - v\|^2/\tau^2$ et $g_k \equiv 1$, pour tout $k \in [K]$. Sous ces conditions et restrictions, nous établissons

la normalité asymptotique de l'estimateur (5). Les détails peuvent être trouvés dans la Section 5 de Gómez-García *et al.* (2020) et seront aussi discutés lors de cette présentation.

Références

Doukhan, P. and Wintenberger, O. (2008) *Weakly dependent chains with infinite memory.* Stochastic Processes and their Applications, 118 :1997–2013.

Gómez-García, JG., Fadili, J. and Chesneau, C. (2020) *Learning CHARME models with (deep) neural networks.* arxiv preprint arxiv :2002.03237.

Kirch, C. and Kamgaing, T. (2012) *Testing for parameter stability in nonlinear autoregressive models.* Journal of Time Series Analysis, 33(3) :365–385.

Liehr, S., Pawelzik, K., Kohlmorgen, J. and Moler, K.R. (1999) *Hidden markov mixtures of experts with an application to eeg recordings from sleep.* Th. of Biosc, 118 :246–260.

Lo, M.T., Tsai, P.H., Lin, P.F., Lin, C. and Hsin, Y.L. (2009) *The nonlinear and nonstationary properties in eeg signals : probing the complex fluctuations by hilbert-huang transform.* Advances in Adaptive Data Analysis, 1(3) :461–482.

Rockafellar, R.T. (1976) *Integral functionals, normal integrands and measurable selections.* In J. Gossez and L. Waelbroeck, editors, *Nonlinear Operators nd the Calculus of Variations*, number 543 in Lecture Notes in Mathematics, pages 157–207. Springer.

Rockafellar, R.T. and Wets, R.J.B. (1998) *Variational Analysis.* Springer.

Stockis, J-P., Franke, J. and Tadjuidje Kamgaing, J. (2010) *On geometric ergodicity of charme models.* Journal of Time Series Analysis, 31 :141–152.

Tadjuidje-Kamgaing, J. (2005) *Competing neural networks as model for nonstationary financial time series.* PhD thesis, University of Kaiserslautern.

Weigend, A.S. and Shi, S. (2000) *Predicting daily probability distributions of s&p500 returns.* Journal of Forecasting, 19(4) :375–392.

Yarotsky, D. (2017) *Error bounds for approximations with deep relu networks.* Neural Networks, 94 :103–114, 2017.

A Kernel-based Consensual Aggregation for Regression

Sothea HAS

LPSM, Sorbonne Université
75005 Paris, France
sothea.has@lpsm.paris

Résumé. Dans cet exposé, nous introduisons une méthode d'agrégation consensuelle basée sur un noyau pour les problèmes de régression. Nous visons à combiner de manière flexible des régresseurs $r_1, r_2, ..., r_M$ en utilisant une moyenne pondérée où les poids sont définis à l'aide d'une fonction noyau. Cette méthode peut être considérée comme une méthode à noyau standard mise en œuvre sur les prédictions données par tous les estimateurs individuels au lieu des entrées. Ce travail étend le contexte de Biau et al. (2016) à un cadre plus général à l'aide d'une structure de noyau. Nous montrons que cette configuration hérite asymptotiquement de la propriété de consistance des estimateurs consistants. De plus, nous proposons d'apprendre numériquement le paramètre de la méthode en utilisant un algorithme de descente de gradient pour des choix appropriés de fonction noyau au lieu d'utiliser l'algorithme de recherche de grille classique. Les expériences numériques réalisées sur plusieurs jeux de données simulés et réels suggèrent que la performance de la méthode est améliorée avec l'introduction des fonctions noyau.

Mots-clés. Agrégation consensuelle, noyau, régression.

Abstract. In this talk, we introduce a kernel-based consensual aggregation method for regression problems. We aim to flexibly combine individual regression estimators $r_1, r_2, ..., r_M$ using a weighted average where the weights are defined based on some kernel function. It may be seen as a kernel smoother method implemented on the features of predictions, given by all the individual estimators, instead of the original inputs. This work extends the context of Biau et al. (2016) to a more general kernel-based framework. We show that this configuration asymptotically inherits the consistency property of the basic consistent estimators. Moreover, we propose to numerically learn the key parameter of the method using a gradient descent algorithm for a suitable choice of kernel functions instead of using the classical grid search algorithm. The numerical experiments carried out on several simulated and real datasets suggest that the performance of the method is improved with the introduction of kernel functions.

Keywords. Consensual aggregation, kernel, regression.

1 Introduction

The first idea of consensual aggregation method was introduced in classification problem by Mojirsheibani (1999). In his combing classification method, the predicted class of a given data point is the majority vote among the actual classes of those data points that share the same predictions, given by all the individual classifiers, with the query point. Later, Mojirsheibani (2000) and Mojirsheibani and Kong (2016) introduced respectively the exponential and general kernel-based versions of the primal method.

Analogously, Biau et al. (2016) configured the primary idea of Mojirsheibani (1999) as regression framework where a training point x_i is "close" to the query point x if each of their predictions, given by all the basic regression estimators, is close. The combination is the weighted average of the actual response values y_i of the close neighbors x_i of x. In this case, the weight is defined using 0-1 loss. It was shown theoretically in these former papers that the combinations inherit the consistency property of consistent basic estimators.

Recently, a kernel-based version of Biau et al. (2016) called `KernelCobra` has been implemented in `pycobra` python library (see Guedj and Srinivasa Desikan (2018)). Moreover, it has also been applied in filtering to improve the image denoising (see Guedj and Rengot (2020)). In a slightly different setting, we present another kernel-based consensual regression aggregation method in this paper, as well as its theoretical and numerical performances. We show that the consistency inheritance property shown in Biau et al. (2016) also holds for this kernel-based configuration for a broad class of regular kernels. Moreover, the evidence of numerical simulation carried out on a similar set of simulated models, and some real datasets show that the present method outperforms the classical one.

2 The proposed kernel-based method

2.1 Notation and definition

We consider a training sample $\mathcal{D}_n = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$ where $(X_i, Y_i), i = 1, 2, ..., n$, are independent and identically distributed with the same realization as (X, Y). We assume that (X, Y) is an $\mathbb{R}^d \times \mathbb{R}$-valued random variable with a suitable integrability which will be specified later. We randomly split $\mathcal{D}_n$ into two subsets $\mathcal{D}_k = \{(X_1^{(k)}, Y_1^{(k)}), (X_2^{(k)}, Y_2^{(k)}), ..., (X_k^{(k)}, Y_k^{(k)})\}$ and $\mathcal{D}_\ell = \{(X_1^{(\ell)}, Y_1^{(\ell)}), (X_2^{(\ell)}, Y_2^{(\ell)}), ..., (X_\ell^{(\ell)}, Y_\ell^{(\ell)})\}$ of size k and ℓ respectively such that $k + \ell = n$ (the common choice is $k = \lceil n/2 \rceil = n - \ell$). The M individual candidate regression estimators, $r_{k,1}, r_{k,2}, ..., r_{k,M}$, are constructed using only the data points contained in $\mathcal{D}_k$ and the choice of these regression estimators are arbitrary. They could be parametric, nonparametric or semi-parametric, and the only requirement of these individual estimators is being able to provide predictions of the remaining part $\mathcal{D}_\ell$ and the query point x. In the sequel, for any $x \in \mathbb{R}^d$, the following notations are used:

- $\mathbf{r}_k(x) = (r_{k,1}(x), r_{k,2}(x), ..., r_{k,M}(x))$: the vector of predictions of x.
- $\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$: Euclidean norm on $\mathbb{R}^d$.
- $g^*(x) = \mathbb{E}[Y|X = x]$: the optimal regression estimator.

A regular kernel, $K : \mathbb{R}^M \to \mathbb{R}_+$, is a nonnegative decreasing function satisfying:

$$\exists b, \kappa_0, \rho > 0 \text{ such that } \begin{cases} b\mathbb{1}_{B_M(0,\rho)}(z) \leq K(z) \leq 1, \forall z \in \mathbb{R}^M \\ \int_{\mathbb{R}^M} \sup_{u \in B_M(z,\rho)} K(u)dz = \kappa_0 < +\infty \end{cases} \tag{1}$$

where $B_M(c, r) = \{z \in \mathbb{R}^M : \|c - z\| < r\}$ denotes the open ball of center $c \in \mathbb{R}^M$ and radius $r > 0$. The consensual regression aggregation evaluated at any point $x \in \mathbb{R}^d$ is defined by,

$$g_n(\mathbf{r}_k(x)) = \sum_{i=1}^{\ell} W_{n,i}(x) Y_i^{(\ell)}. \tag{2}$$

The proposed method in this talk corresponds to the following weight:

$$W_{n,i}(x) = \frac{K_h(\mathbf{r}_k(X_i^{(\ell)}) - \mathbf{r}_k(x))}{\sum_{j=1}^{\ell} K_h(\mathbf{r}_k(X_j^{(\ell)}) - \mathbf{r}_k(x))}, i = 1, 2, ..., \ell \tag{3}$$

where $K_h(z) = K(z/h)$ for some smoothing parameter $h > 0$ with the convention of $0/0 = 0$. Notice that the classical method of Biau et al. (2016) corresponds to the following naive weight:

$$W_{n,i}(x) = \frac{\prod_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(X_i^{(\ell)}) - r_{k,m}(x)| < \varepsilon\}}}{\sum_{j=1}^{\ell} \prod_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(X_j^{(\ell)}) - r_{k,m}(x)| < \varepsilon\}}}, i = 1, 2, ..., \ell \tag{4}$$

for some smoothing parameter $\varepsilon > 0$ with the same convention of $0/0 = 0$.

2.2 Theoretical performance

The performance of the combining estimation g_n is measured using the following quadratic risk,

$$\mathbb{E}\Big[|g_n(\mathbf{r}_k(X)) - g^*(X)|^2\Big]$$

where the expectation is taken with respect to both X and the training sample $\mathcal{D}_n$. The following proposition shows that the nonasymptotic-type control of the distortion, presented in Proposition.2.1 of Biau et al. (2016), also holds for this case of regular kernels.

Proposition 1 *Let $\boldsymbol{r}_k = (r_{k,1}, r_{k,2}, ..., r_{k,M})$ be the collection of all basic estimators and $g_n(\boldsymbol{r}_k(x))$ be the combined estimator computed at point $x \in \mathbb{R}^d$. Then, for all distributions of (X, Y) with $\mathbb{E}[|Y|^2] < +\infty$,*

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f \in \mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] + \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big].$$

In particular,

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \min_{1 \leq m \leq M} \mathbb{E}\Big[|r_{k,m}(X) - g^*(X)|^2\Big] + \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big].$$

Given all the machines, the first term of this bound cannot be controlled as it depends on the performance of the best constructed machine, and it will be there as the asymptotic control of the performance of the proposed method. Our main task is to deal with the second term $\mathbb{E}[|g_n(\mathbf{r}_k(X)) - g^*(\mathbf{r}_k(X))|^2]$, which can be shown to be asymptotically negligible in the following key proposition.

Proposition 2 *Assume that $r_{k,m}$ is bounded for all $m = 1, 2, .., M$. Let $h \to 0$ and $\ell \to +\infty$ such that $h^M \ell \to +\infty$. Then*

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \to 0 \text{ as } \ell \to +\infty$$

for all distribution of (X, Y) with $\mathbb{E}[|Y|^2] < +\infty$. Thus,

$$\limsup_{\ell \to +\infty} \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f \in \mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

And in particular,

$$\limsup_{\ell \to +\infty} \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \min_{1 \leq m \leq M} \mathbb{E}\Big[|r_{k,m}(X) - g^*(X)|^2\Big].$$

Proposition 2 above is an analogous setup of Proposition 2.2 in Biau et al. (2016). In this study, we can derive the result for this broader class thanks to the boundedness of all the basic machines. However, the price to pay for the universality for this class of regular kernels is the lack of convergence rate. To this goal, a weak smoothness assumption of g^* with respect to the basic machines is required. For example, the convergence rate obtained in Biau et al. (2016) is of order $O(\ell^{-2/(M+2)})$ under the same smoothness assumption, and this result holds for all the compactly support kernels. Our goal is not to theoretically do better than the classical method but to investigate such a similar question in a broader class of kernel functions. For those kernels which the tails decrease fast enough, the convergence rate of the variance-type term can be attained as described in the following main theorem of this talk.

Theorem 1 *Assume that the response variable Y and all the basic machines $r_{k,m}, m = 1, 2, ..., M$, are bounded by some constant R. Suppose that there exists a constant $L \geq 0$ such that, for every $k \geq 1$,*

$$|g^*(\boldsymbol{r}_k(x)) - g^*(\boldsymbol{r}_k(y))| \leq L\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(y)\|, \forall x, y \in \mathbb{R}^d.$$

We assume moreover that,

$$\exists R_K, C_k > 0 : K(z)\|z\|^2 \leq \frac{C_K}{1 + \|z\|^M}, \forall z \in \mathbb{R}^M \text{ such that } \|z\| \geq R_K.$$

Then, with the choice of $h \propto \ell^{-\frac{M+2}{M^2+2M+4}}$, one has

$$\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2] \leq \min_{1 \leq m \leq M} \mathbb{E}[|r_{k,m}(X) - g^*(X)|^2] + C\ell^{-\frac{4}{M^2+2M+4}} \tag{5}$$

for some positive constant $C = C(b, L, R, R_K, C_K)$ independent of ℓ.

Remark 1 *The assumption on the upper bound of the kernel K in the theorem above is very weak, chosen so that the result holds for a large subclass of regular kernels. However, the convergence rate is indeed slow for this subclass of kernel functions. If we strengthen this condition, we can obtain a much nicer result. For instance, if we assume that the tails decrease exponentially fast i.e.,*

$$\exists R_K, C_K > 0 \text{ and } \alpha \in (0, 1) : K(z) \leq C_K e^{-\|z\|^\alpha}, \forall z \in \mathbb{R}^M, \|z\| \geq R_K,$$

by following the same procedure as in the proof of the above theorem, one can easily check that the convergence rate (of the quantity in proposition 2) is of order $O(\ell^{-2\alpha/(M+2\alpha)})$. This rate approaches the state of the art of the classical method by Biau et al. (2016) when α approaches 1.

3 A summary of numerical result and conclusion

Constructing the proposed method is equivalent to estimating the window parameter which was normally done in all the previous studies using grid search algorithm. Through several simulations, we observe a convex-like curve of the quadratic risk (as function of window parameter). To take benefit from this observation, we propose to estimate the key parameter using gradient descent algorithm for suitable options of kernel functions such as Gaussian kernel, for example. Several numerical experiments carried out on different simulated datasets (see Biau et al. (2016)), real public datasets (Dua and Graff (2017a), Kaggle (2016) and Cortez et al. (2009), Dua and Graff (2017b)) and private dataset (Cadet et al. (2005)), confirm the theoretical results described in the previous section. We observe that the average errors produced by the proposed method mostly outperform or bias towards the best individual estimator of the combination. Moreover, the performance is improved compared to the classical method by Biau et al. (2016).

References

Biau, G., Fischer, A., Guedj, B., Malley, J.D., 2016. COBRA: a combined regression strategy. Journal of Multivariate Analysis 146, 18–28.

Cadet, O., Harper, C., Mougeot, M., 2005. Monitoring energy performance of compressors with an innovative auto-adaptive approach., in: Instrumentation System and Automation -ISA- Chicago.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis., J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, Elsevier 47, 547–553.

Devroye, L., Györfi, L., Lugosi, G., 1997. A Probabilistic Theory of Pattern Recognition. Springer.

Dua, D., Graff, C., 2017a. UCI machine learning repository: Abalone data set.

Dua, D., Graff, C., 2017b. UCI machine learning repository: Wine quality data set.

Fischer, A., Mougeot, M., 2019. Aggregation using input-output trade-off. Journal of Statistical Planning and Inference 200, 1–19.

Guedj, B., Rengot, J., 2020. Non-linear aggregation of filters to improve image denoising, in: Arai, K., Kapoor, S., Bhatia, R. (Eds.), Intelligent Computing, Springer International Publishing, Cham. pp. 314–327.

Guedj, B., Srinivasa Desikan, B., 2018. Pycobra: A python toolbox for ensemble learning and visualisation. Journal of Machine Learning Research 18, 1–5.

Guedj, B., Srinivasa Desikan, B., 2020. Kernel-based ensemble learning in python. Information 11, 63. doi:10.3390/info11020063.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer.

Kaggle, 2016. House sales in king county, usa.

Mojirsheibani, M., 1999. Combined classifiers via disretization. Journal of the American Statistical Association 94, 600–609.

Mojirsheibani, M., 2000. A kernel-based combined classification rule. Journal of Statistics and Probability Letters 48, 411–419.

Mojirsheibani, M., Kong, J., 2016. An asymptotically optimal kernel combined classifier. Journal of Statistics and Probability Letters 119, 91–100.

Prise en compte de la structure temporelle dans l'analyse de données protéomiques à haut débit

Wilfried Heyse [1], Vincent Vandewalle [2], Philippe Amouyel [3], Guillemette Marot [4], Christophe Bauters [5] & Florence Pinet [6].

[1] *U1167 (Inserm, Université de Lille, CHU Lille, Institut Pasteur de Lille) & Équipe Projet Modal (Inria Lille - Nord Europe), wilfried.heyse@inria.fr*

[2] *ULR2694 (Université de Lille) & Équipe Projet Modal (Inria Lille - Nord Europe), vincent.vandewalle@inria.fr*

[3] *U1167 (Inserm, Université de Lille, CHU Lille, Institut Pasteur de Lille) philippe.amouyel@pasteur-lille.fr*

[4] *ULR2694 (Université de Lille) & Équipe Projet Modal (Inria Lille - Nord Europe), guillemette.marot@univ-lille.fr*

[5] *U1167 (Inserm, Université de Lille, CHU Lille, Institut Pasteur de Lille) christophe.bauters@chru-lille.fr*

[6] *U1167 (Inserm, Université de Lille, CHU Lille, Institut Pasteur de Lille) florence.pinet@pasteur-lille.fr*

Résumé. Chaque année, en France, plus de 100 000 personnes déclarent un infarctus du myocarde (IM) qui, pour certains d'entre eux, conduit à un remodelage ventriculaire gauche (RVG) et à de l'insuffisance cardiaque (IC). De précédentes études ont montré que la présence d'un RVG suite à un infarctus était un facteur de risque d'IC et de décès pour causes cardiovasculaires. La rechercher de biomarqueurs permettant la prédiction du RVG ou de la survie à un stade précoce est donc un problème de santé publique. Notre but, ici, est de sélectionner un petit nombre de protéines liées au RVG ou à la survie en utilisant les mesures de plus de 5000 protéines sur deux cohortes d'environs 240 patients chacune disponibles au moment de l'infarctus, mais aussi à trois temps supplémentaires pour l'une des deux cohortes. Dans un premier temps, nous présenterons un modèle prédictif de la survie basé sur la création de clusters de patients. Puis, nous nous concentrerons sur la dimension longitudinale des données et explorerons comment cette dimension peut nous être utile dans la sélection de protéines pour la prédiction précoce de la survie des patients. Afin de modéliser la dimension longitudinale et la grande dimension des données un clustering longitudinal sera d'abord étudié afin de créer des groupes de protéines pouvant ensuite être utilisés dans un modèle de prédiction de la survie.

Mots-clés. Santé, prédiction, clustering, survie, données longitudinales, grande dimension.

Abstract. Each year, in France, over 100 000 people suffer from myocardial infarction (MI) which, for some of them, lead to a left ventricular remodeling (LVR) and heart failure (HF). Studies have shown that during a year following MI, LVR is a risk factor for HF and cardiovascular death. Finding biomarkers which can detect early stage of

LVR or HF after a MI is a leading public health matter. We are aiming at selecting few proteins responsible for LVR and survival, using not only baseline measurements of over 5000 proteins on 2 cohorts of around 240 patients each, but also using three additional longitudinal measurements of these proteins available on one of the two cohorts. In a first time, we will present how we developped a prediction survival model by creating cluster of patients. In a second time, we will focus on the longitudinal dimension of the data and explore how this dimension could help selecting relevant proteins for predicting survival using only baseline measurment. To handle the longitudinal (and high) dimension of the data, clustering of longitudinal data will be studied in order to create groups of proteins that could be used in a selection model.

Keywords. Health, prediction, clustering, survival, longitudinal data, high dimension.

1 Introduction

1.1 Contexte clinique

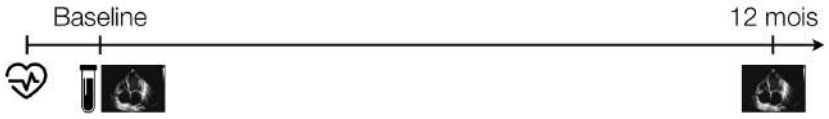
Dans un contexte médical où plus de 70 000 personnes meurent chaque année en France des suites d'insuffisance cardiaque chronique, il est devenu très important de comprendre les causes de ce phénomène et de le prévenir. L'insuffisance cardiaque est un état de santé global et complexe d'un patient qui peut résulter de multiples causes et qui se traduit par une incapacité du cœur à pomper suffisamment de sang pour assurer à l'ensemble des organes du corps humain un apport en oxygène suffisant pour fonctionner correctement. La maladie coronarienne et en particulier l'infarctus du myocarde est la cause la plus fréquente d'insuffisance cardiaque. En effet, les parties du cœur touchées par la nécrose vont totalement perdre leur fonction de contraction rendant la pompe cardiaque moins efficiente.

Suite à un infarctus du myocarde, un second phénomène délétère va parfois se mettre en place, il s'agit du Remodelage Ventriculaire Gauche (RVG). Le RVG est une dilatation progressive du ventricule gauche. En effet, les tissus nécrosés, qui ont perdu leur fonction de contraction vont progressivement se relâcher. Ce phénomène complexe de RVG, qui va conduire le cœur à constamment perdre de la fonction, a été identifié comme un indicateur puissant d'un risque élevé d'insuffisance cardiaque ou de décès cardiovasculaire après un infarctus du myocarde.

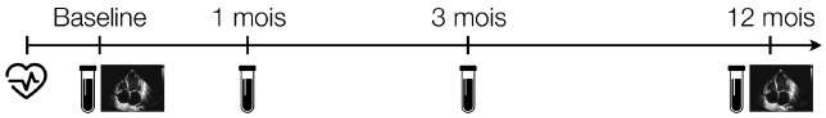
1.2 Cohortes REVE

Les cohortes REVE (REmodelage VEntriculaires), coordonnées par le Pr. Bauters, ont été conçues spécifiquement afin d'étudier le RVG après un premier infarctus. L'étude

REVE-1 a inclus 266 patients entre 2002 et 2004 et l'étude REVE-2 a inclus 246 patients entre 2006 et 2008. Les patients des deux études présentent des infarctus du myocarde du même type dans le sens où tous les patients inclus ont eu un infarctus causé par une obstruction de l'artère coronaire interventriculaire antérieure et pour laquelle l'obstruction est survenue au niveau proximal, c'est-à-dire au début de l'artère. Ce critère a été choisi pour sélectionner des patients qui ont un infarctus massif permettant d'observer plus de patients avec un RVG important.



(a) Cohorte REVE-1



(b) Cohorte REVE-2

FIGURE 1 – Cohortes REVE

Les patients inclus dans ces études ont été suivis durant 1 an afin de quantifier leur RVG (Figure 1). Afin d'identifer des marqueurs prédictifs du RVG, les patients de l'étude REVE-1 ont eu un prélèvement sanguin durant leur hospitalisation (prélèvement en baseline). Les patients de l'étude REVE-2 ont eu 4 prélèvements sériés tout au long de l'année de suivi. Les prélèvements sanguins effectués lors des deux études ont été prélevés selon un protocole très strict afin d'éviter toute variabilité due aux protocoles de prélèvements sanguins. Sur ces deux cohortes, un suivi à long terme des patients a été effectué permettant d'identifier les patients ayant été hospitalisé pour insuffisance cardiaque ou étant décédés d'une cause cardiovasculaire, celui-ci a permis de montrer les liens entre RVG important et survie [1]. Nous nous concentrerons ici sur l'exploitation des données produites par ces études afin de prédire la survie à long terme des patients.

2 Analyse de la survie par clustering

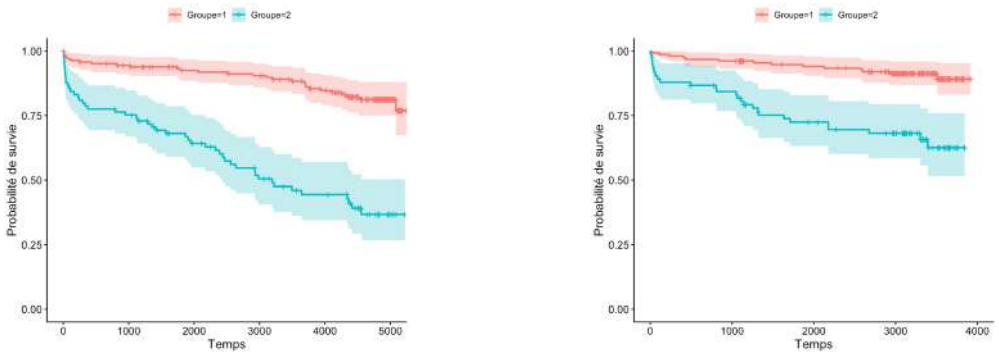
Les prélèvements sanguins obtenus lors des deux études ont permis grâce à une collaboration avec la société SomaLogic de quantifier la concentration dans le sang de 5284 protéines (cette techinique très innovante permet à ce jour la plus grande quantification simultanée de protéines). Nous avons donc à disposition les mesures de 5284 protéines pour 266 patients (REVE-1) au temps baseline et pour 246 patients (REVE-2) à 4 temps

différents.

Une première analyse, sans prise en compte de la structure temporelle nous à permis de déterminer des clusters de patients basés uniquement sur les mesures protéomiques qui ont un fort pouvoir prédictif sur la survie à long terme. Le C-index [2] [3] a été utilisé afin de mesurer la qualité prédictive des modèles présentés. Nous nous comparons à un modèle clinique basé uniquement sur les variables cliniques. Ce modèle, appris sur REVE-1 puis validé sur REVE-2 a montré de bonnes performances avec un C-index de 0.75 sur la cohorte d'apprentissage et de 0.77 sur la cohorte de validation.

Nous avons d'abord pré-sélectionné les protéines qui étaient significativement associées à la survie après correction des p-valeurs par Bonferroni (50 protéines ont ainsi été sélectionnées). Nous avons créé des groupes de patients grâce à un k-means. Le nombre de clusters permettant d'optimiser la prédiction de la survie a sélectionné 2 clusters.

L'information seule de l'appartenance à l'un ou l'autre des clusters a permis d'augmenter les performances prédictives du modèle de survie en terme de C-index passant à 0.83 dans la cohorte d'apprentissage et 0.82 dans la cohorte de validation et permettant de très bien discriminer les patients selon leur survie comme le montre le figure 2. Ces résultats se révèlent être meilleurs que les résultats obtenus avec les méthodes de sélection de variables en grande dimension (type LASSO) sur ces données. Cela nous a permis d'établir l'intérêt de l'étude des données protéomiques dans la prédiction de la survie.



(a) Cohorte REVE-1 (Apprentissage) (b) Cohorte REVE-2 (Validation)

FIGURE 2 – Survie des patients selon le cluster attribué

3 Clustering de données longitudinales

La suite de notre travail consiste à exploiter les 4 prélèvements de l'étude REVE-2. Cette étude a deux buts, le premier est de permettre une meilleure compréhension de

l'évolution temporelle des protéines, le second est de permettre une meilleure sélection des protéines afin de les utiliser dans la prédiction de la survie à long terme.

Afin d'exploiter la structure temporelle et de répondre à la première question, nous avons décidé de créer des groupes de protéines qui varient de la même manière au cours du temps afin de réduire la dimension du problème. En choisissant un nombre de groupes G, nous souhaitons donc passer de données de taille $n \times p \times T$ à des données de taille $n \times G \times T$ avec $G \ll p$.

Notons x_{ijt} l'expression de la protéine j pour l'individu i, au temps t et $\mathbf{x}_j = (x_{ijt})_{i,t}$: ensemble des mesures de la protéine j représentées sous forme d'une matrice $n \times T$. Afin créer des clusters de protéines, un modèle de mélange [4] est ajusté sur les $\mathbf{x}_j$:

$$g(\mathbf{x}_j) = \sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_j) \tag{1}$$

où g est la loi du modèle de mélange permettant de modéliser les protéines, π_k et f_k respectivement les proportions et densités dans la classe k.

Ainsi, pour chaque protéine $\mathbf{x}_j$, nous avons :
— $\mathbf{z}_j = (z_{j1}, ..., z_{jG}) \sim \mathcal{M}(\pi_1, ..., \pi_G)$ la classe de la protéine $\mathbf{x}_j$
— $\mathbf{x}_j | z_{jk} = 1 \sim MM(\theta_k)$, sachant sa classe, la protéine $\mathbf{x}_j$ est modélisée par le modèle linéaire mixte :

$$x_{ijt} = \mu_k + b_{ij} + \beta_{kt} + \varepsilon_{ijt} \tag{2}$$

Avec :
— μ_k l'effet fixe des protéines de la classe k
— $b_{ij} | z_{jk} = 1 \sim \mathcal{N}(0, \sigma^2_{1,k})$ l'effet aléatoire de l'individu i pour les protéines de la classe k
— β_{kt} l'effet temporel fixe t pour les protéines de la classe k
— $\varepsilon_{ijt} | z_{jk} = 1 \sim \mathcal{N}(0, \sigma^2_{2,k})$ le terme d'erreur pour les protéines de la classe k.

Grâce à un tel modèle, nous pouvons créer des clusters de protéines dont l'evolution temporelle est proche, comme montré sur la figure 3.

Cette modélisation nous permet une certaine flexibilité dans la modélisation des clusters, en effet le modèle linéaire mixte décrit par l'équation 2 permet de modéliser finement les effets à prendre en compte, mais aussi d'ajouter des effets cliniques des patients afin d'ajuster les clusters sur un certain nombre de paramètres cliniques.

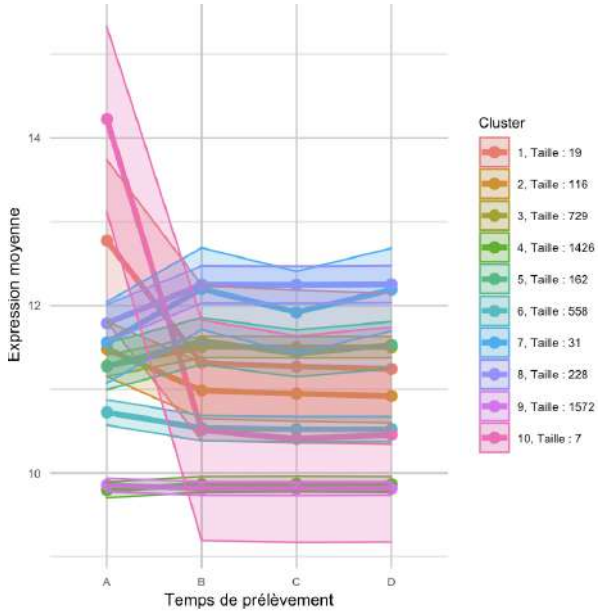


FIGURE 3 – Profil moyen des différents clusters créés

Les perspectives de ce travail sont donc d'exploiter l'information des groupes de protéines dans une méthode de sélection de variables afin d'améliorer la pertinence et l'interprétatbilité des protéines sélectionnées et d'accroître les performances du modèle de prédiction de la survie.

Références

[1] Christophe Bauters, Emilie Dubois, Sina Porouchani, Eric Saloux, Marie Fertin, Pascal de Groote, Nicolas Lamblin, and Florence Pinet. Long-term prognostic impact of left ventricular remodeling after a first myocardial infarction in modern clinical practice. *PLOS ONE*, 12(11) :1–13, 11 2017.

[2] Frank Harrell, Kerry Lee, and Daniel Mark. Multivariable prognostic models : Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4) :361–387, feb 1996.

[3] Michael Pencina and Ralph D'Agostino. Overall c as a measure of discrimination in survival analysis : model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13) :2109–2123, 2004.

[4] Gilles Celeux, Christian Lavergne, and Olivier Martin. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5 :243–267, November 2005.

Prendre en compte la variabilité spatio-temporelle du micro habitat climatique dans un modèle mécanistique de répartition d'espèce ; un défi à la hauteur de différentes méthodes statistiques.

Florèn Hugon [1], Claire Kermorvant [1], Aurélien Callens [1], Bastien Mourguiart [1], Sébastien Coube [1], Teo Nguyen [1], Benoit Liquet [1,2] & Frank D'Amico [1]

[1] *CNRS/Univ Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications de Pau - Fédération MIRA, UMR 5142, 64600 Anglet, France; floren.hugon@univ-pau.fr, claire.kermorvant@univ-pau.fr, aurelien.callens@univ-pau.fr, bastien.mourguiart@univ-pau.fr, sebastien.coube@univ-pau.fr, teo.nguyen@univ-pau.fr, benoit.liquet@univ-pau.fr, frank.damico@univ-pau.fr*

[2] *Department of Mathematics and Statistics, Macquarie University, Australia*

Résumé. Communément, lorsque des études s'intéressent aux effets du changement climatique sur la biodiversité, seules les modifications moyenne des variables climatiques sont étudiées. Or, de nombreuses études montrent le rôle, tout aussi important, de la variabilité sur la mise en place des réponses des espèces face au changement climatique. Les espèces ectothermes, qui régulent leur température corporelle en modifiant leur comportement, sont particulièrement sensibles au changement climatique. Le temps d'activité au cours de la période critique de la reproduction, défini par les capacités physiologiques d'une espèce, est une variable indicatrice de la persistance. Établir un modèle de projection du temps d'activité en fonction de variables climatiques permettrait de projeter la répartition des espèces ectothermes. Ici, des modèles biomimétiques pour le lézard de Bonnal (*Iberolacerta bonnali*) ont été déployés, au cours de sa saison de reproduction, sur différents sites et plusieurs années. Ils ont enregistré un proxy de la température corporelle des individus, utilisé pour calculer le temps d'activité. Nous explorons l'utilisation des Generalized Additive Models (GAMs), des processus Gaussien et de la modélisation bayésienne, pour projeter le temps d'activité en fonction de la température, l'humidité, les précipitations et le vent. Le challenge de cette étude réside en l'intégration de la variabilité spatio-temporelle des séries de temps d'activité, dans le modèle de projection. Les meilleurs modèles issus des meilleures méthodes seront utilisés pour projeter la répartition de l'espèce aux horizons 2050, 2070 et 2100 sous les scénarios RCP 2.6, 4.5 et 8.5.

Mots-clés. Modèles Additifs Généralisés, Modèles Mixtes, Modélisation Bayésienne, Processus Gaussien, Projection de Répartition, Temps d'Activité.

Abstract. Commonly, when studies focus on the climate change effects on biodiversity, only mean changes in climate variables are studied. However, many studies show the equally important role of variability in implementation of species responses to climate

change. Ectotherms species, which regulate their body temperature by modifying their behaviour, are particularly sensitive to climate change. The activity time during the critical breeding period, defined by the physiological capacities of a species, is an indicator variable for the persistence. Building a model for projecting activity time as a function of climatic variables would permit to project the distribution of ectothermal species. Here, biomimetic models for the Pyrenean Rock lizard (*Iberolacerta bonnali*) were deployed during its breeding season, on different sites and over several years. They recorded body temperature proxy of individuals, used to calculate activity time. We are exploring the use of Generalized Additive Models (GAMs), Gaussian processes and Bayesian modelling to project activity time as a function of temperature, humidity, precipitation and wind. The study challenge lies in the integration into the projection model of the activity time series spatio-temporal variability. The best models from the best methods will be used to project the species distribution over the 2050, 2070 and 2100 horizons under RCP scenarios 2.6, 4.5 and 8.5.

Keywords. Activity Time, Additive Generalized Models, Bayesian Modelling, Distribution Prediction, Gaussian Process, Mixture Models.

1 Introduction

1.1 Changement climatique et variabilité spatio-temporelle

Communément, lorsque les études traitent des effets statistiques du changement climatique sur la biodiversité, la façon courante de les aborder consiste à se fonder uniquement sur la moyenne, alors que le changement climatique entraîne aussi une plus grande variabilité des paramètres climatiques (Rummukainen 2012). La littérature sur l'impact du changement climatique sur la biodiversité n'aborde la question qu'en terme de moyenne et ignore la variabilité (Estay et *al.* 2011). De nombreuses études reconnaissent la variabilité spatiale et utilisent la réplication spatiale mais finissent par faire la moyenne des résultats des différents sites d'étude (Sinervo et *al.* 2010). Très peu abordent la variabilité temporelle, les expériences sont souvent menées sur une seule année (Arribas 2009). L'absence de véritable réplication spatio-temporelle des études soulève la question cruciale de la représentativité des résultats. En effet, des études démontrent le rôle important de la variabilité dans la mise en place des réponses des espèces au changement climatique (Cahill et *al.* 2012). Celles-ci semblent être dirigées par les régimes thermiques plus que par la température moyenne (Tourneur et *al.* 2019). De plus, la moyenne et la variabilité interagissent, pour expliquer par exemple les performances thermiques des espèces ectothermes (Bozinovic et *al.* 2011). Ainsi, la variabilité climatique semble jouer un rôle majeur dans la persistance des espèces, il est donc crucial de l'intégrer dans les études traitant des effets du changement climatique sur la biodiversité (Clusella-Trullas et *al.* 2011).

1.2 Le temps d'activité, une variable mécanistique utilisée pour projeter la répartition d'une espèce

Les espèces ectothermes, comme les reptiles, contrôlent leur température corporelle par leur comportement et leur physiologie (Theisinger 2016). Leurs performances métaboliques sont dépendantes de la température, ce qui conduit au comportement de thermorégulation (Sinclair et *al.* 2016). Afin de maintenir leur température corporelle dans une gamme qui optimise la performance des processus physiologiques, ces espèces alternent entre des périodes d'activités dynamiques (chasse, reproduction), des périodes de chauffe (lézardage) et des périodes en refuge, leur permettant d'éviter des températures extérieures trop fraîches ou trop chaudes (Arribas 2009, Huey et *al.* 2009). Face au changement climatique, la réponse la plus observée est le changement du pattern d'activité quotidien (Theisinger 2016). Cette réponse est efficace mais pourrait devenir inadéquate à long terme, si elle entraîne une réduction trop importante du temps d'activité (Kearney 2013). Cela conduirait à une baisse du succès reproducteur (Ortega et *al.* 2016) et à la hausse du risque d'extinction locale (Cahill et *al.* 2012). Ainsi, le temps d'activité pendant la période critique de reproduction peut être utilisé comme un indicateur de la persistance des populations et permettre la projection de la répartition (Ceia-Hasse et *al.* 2014).

2 Matériels et Méthodes

2.1 Des températures opérantes au calcul du temps d'activité, variable explicative de la répartition

Le Lézard de Bonnal, (*Iberolacerta bonnali*) est une espèce endémique des Pyrénées, inféodée à l'étage alpin (Pottier 2012). Afin d'étudier la variabilité spatio-temporelle de l'environnement climatique de cette espèce, des modèles biomimétiques, qui enregistrent un proxy de la température corporelle, appelée température opérante (Te), ont été déployés sur trois sites d'étude de 2017 à 2020 (Hugon et *al.* 2020). Les séries sont enregistrées sur la période de reproduction à une fréquence de 10 minutes. La variabilité intra-site est étudiée selon l'altitude (2200 mètres *versus* 2400) et le versant (nord *versus* sud) par le suivi sur le site de Peyreget où quatre localités sont définies, mais aussi par le déploiement de réplicats sur Anglas et Arrious. Les suivis en 2019 à une altitude similaire pour les trois sites permettent l'étude de la variabilité inter-site. Par hiérarchie, pour une même année, nous pouvons distinguer dans les modèles, un niveau site puis un niveau intra-site (réplicat ou localité). Les suivis pluri-annuels sur Anglas et Peyreget permettent l'étude de la variabilité annuelle. Pour chacune des 25 séries de températures opérantes, les séries de temps d'activité journalier et horaire ont été calculées. Le temps d'activité correspond à la somme des périodes de 10 minutes où Te est incluse dans la fenêtre thermique d'activité, bornée par les températures volontaires d'activité, $VTmin$=20.8C et $VTmax$=35.2C (Caetano et al. 2020) La série obtenue permet de calculer le temps

d'activité total sur la période de reproduction, de déduire l'indice de persistance, pour ensuite projeter la répartition de l'espèce (Équation 1, Hugon et *al.* 2020).

$$HaDaily_j = 10 \times HaCount_j/60$$

où $HaCount_j = \sum_i 1_{\{Te_{ij} \in [VTmin, VTmax]\}}$, avec $1_{\{.\}}$ la fonction indicatrice du temps d'activité et Te_{ij} la $i$ème mesure du jour j. Ainsi, le temps d'activité total sur la période de reproduction est :

$$HaTot = \sum_j HaDaily_j \tag{1}$$

2.2 Projection du temps d'activité à partir des simulations climatiques, exploration de diverses méthodes statistiques

Les simulations climatiques utilisées dans notre étude sont celles issues du modèle CNRM-CM5 / ALADIN63. Les variables températures minimale (Tmin), maximale (Tmax) et moyenne, humidité spécifique, précipitations liquides et solides, vitesse du vent sont fournies à la résolution journalière et à la maille 8 kilomètres (grille SAFRAN). Nous utiliserons aussi les simulations du modèle ADAMONT – CNRM-CM5 / ALADIN53 à la résolution journalière et horaire (si les données sont mises à disposition rapidement) et obtenues par massif et tranches d'altitude de 300 mètres. Ce modèle fournit les variables de températures et de précipitations décrites mais également la hauteur de neige et l'équivalent en eau de la neige, qui renseignent indirectement l'humidité du milieu. Afin d'inclure l'écophysiologie de l'activité dans les variables explicatives, nous proposons la définition de nouvelles variables explicatives dépendantes de $VTmin$ et $VTmax$, telles que $Tmax - VTmax$ et $Tmax - VTmin$. Les corrélations seront étudiées entre toutes les variables explicatives afin de choisir les variables à inclure dans la modélisation en limitant la multicolinéarité. La modélisation à la résolution horaire permettrait de construire des modèles probablement plus explicatifs, intégrant mieux la variabilité du temps d'activité au cours de la journée. Dans cette modélisation, nous inclurons une variable heure et la série horaire des températures de l'air, calculée à partir des $Tmin$ et $Tmax$, plutôt que $Tmin$, $Tmax$ et $Tmoyenne$ (Mallard 2019).

Afin d'intégrer la variabilité observée entre les séries de temps d'activité, il est intéressant de construire des modèles pour chaque série et pour une série moyenne par site, en incluant les données de toutes les années de suivi. Nous explorerons d'abord l'utilisation des modèles additifs généralisés (GAM). Limités par une faible quantité de données qui contient une grande variabilité (dans notre cas d'étude), nous proposerons d'intégrer l'effet site dans un modèle hiérarchique afin d'obtenir un jeu de données d'entraînement plus

important et une meilleure puissance statistique. Également, nous testerons l'utilisation des processus Gaussien en régression (Gramacy 2020). Cette méthodologie permettrait de modéliser la variabilité du temps d'activité, *via* la modélisation du bruit (Bishop 2006, Dorazio 2015), et de l'intégrer dans le modèle de projection. De plus, l'apprentissage à partir de *priors* permet d'outrepasser la limite liée aux données observées au cours de l'entraînement du modèle, présente dans les GAM (Bishop 2006). Cette méthode peut cependant présenter des limites calculatoires, selon la taille du jeu de données et les hyper-paramètres choisis (Gramacy 2020). Enfin, une modélisation bayésienne sera aussi envisagée ; bien qu'elle demandera un investissement plus important, sans promesse de résultat, contrairement aux processus Gaussien (Gramarcy 2020). La définition de *priors* suffisamment informatifs pour permettre la convergence du modèle constituera la principale limite (Dorazio 2015). Cette méthode permettrait de prendre en compte l'incertitude dans la construction et la sélection des modèles en calculant un modèle moyen (Ellison 2004). En plus d'inclure l'incertitude des données, elle reste fiable même pour une faible taille du jeu de données (Dorazio 2015), ce qui constitue un réel atout dans notre cas d'étude.

Le modèle présentant le meilleur pouvoir prédictif sera sélectionné pour projeter le temps d'activité. Cette sélection sera réalisée selon différents critères d'évaluation, le pourcentage de déviance expliquée ainsi que le GCV (Generalized Cross-Validation score) pour les modèles GAM et le BIC (Bayesian Information Criterion) pour les processus Gaussien et les modèles bayésiens. Les projections de temps d'activité seront calculées à partir des simulations climatiques obtenues aux horizons 2050, 2070 et 2100 sous les scénarios RCP 2.6, 4.5 et 8.5 sur l'ensemble des Pyrénées Atlantiques. A partir des temps d'activité, la probabilité de présence sera calculée pour obtenir une carte de répartition.

Bibliographie

Arribas O.J. (2009), Habitat selection, thermoregulation and activity of the Pyrenean Rock Lizard Iberolacerta bonnali (LANTZ, 1927), *Herpetozoa*, vol. 22, num. 3, pp 145-166.

Bishop, C.M. (2006). 6.4 Kernels methods : Gaussian processes, *Pattern recognition and machine learning*, pp. 303-323.

Bozinovic F. et *al.* (2011), The mean and variance of environmental temperature interact to determine physiological tolerance and fitness, *Physiological and Biochemical Zoology*, vol. 4, num. 6, pp 543-552.

Caetano G.H.O et *al.* (2020), Time of activity is a better predictor of the distribution of a tropical lizard than pure environmental temperatures, *Oikos*.

Cahill A.E. et *al.* (2012), How does climate change cause extinction?, *Biological Sciences*, vol. 280, num. 1750.

Ceia-Hasse et *al.* (2014), Integrating ecophysiological models into species distribution projections of European reptile range shifts in response to climate change, *Ecography*, vol. 37, num. 7, pp 679-688.
Clusella-Trullas S. et *al.* (2011), Climatic predictors of temperature performance curve parameters in ectotherms imply complex responses to climate change, vol. 177, num. 6, pp 738-751.
Dorazio, R. M. (2015), Bayesian aata analysis in population ecology: motivations, methods, and benefits, *Population Ecology*, vol. 58, num. 1, pp 31-44.
Ellison A.M. (2004), Bayesian inference in ecology, *Ecology Letters*, vol. 7, num. 6, pp. 509-520
Estay S.A. et *al.* (2011), Beyond average: an experimental test of temperature variability on the population dynamics of Tribolium confusum, *Population Ecology*, vol. 53, num. 1, pp 53-58.
Gramacy, R. B. (2020), Gaussian Process Regression, *Surrogates : Gaussian process modeling, design and optimization for the applied sciences*, pp. 143-222.
Huey R.B. et *al.* (2012), Predicting organismal vulnerability to climate warming: roles of behaviour, physiology and adaptation, *Biological Sciences*, vol. 367, num. 1596, pp 1665-1679.
Hugon, F. et *al.* (2020). Multi-Site and multi-year remote records of operative temperatures with biomimetic loggers reveal spatio-temporal variability in mountain lizard activity and persistence proxy estimates, *Remote Sensing*, vol. 12, num. 18.
Kearney (2013), Activity restriction and the mechanistic basis for extinctions under climate warming, *Ecology Letters*, vol. 16, num.12, pp 1470-1479.
Mallard, F. (2019), Tome VIII : Écologie du changement climatique en région Nouvelle-Aquitaine, Programme les sentinelles du climat, 605p.
Ortega Z. et al. (2016), Behavioral buffering of global warming in a cold-adapted lizard, *Ecology and Evolution*, vol. 6, num. 13, pp 4582-4590.
Pottier G. (2012), Plan national d'actions en faveur des Lézards des Pyrénées Iberolacerta aranica, I. aurelioi et I. bonnali, 2013 - 2017.
Rummukainen M. (2012), Changes in climate and weather extremes in the 21st century: Changes in climate and weather extremes, *Climate Change*, vol. 3, num. 2, pp 115-129.
Sinclair et *al.* (2016), Can we predict ectotherm responses to climate change using thermal performance curves and body temperatures?, *Ecology Letters*, vol. 19, num. 11, pp 1372-1385.
Sinervo B. et *al.* (2010), Erosion of lizard diversity by climate change and altered thermal niches, Science, vol. 328, num. 5980, pp 894-899.
Theisinger (2016), Thermal limits of reptiles. Ecological and environmental constraints on the thermal biology of Malagsy lizards.
Tourneur J-C. et Meunier J.(2019), The successful invasion of the European earwig across North America reflects adaptations to thermal regimes but not mean temperatures, *Evolutionary Biology.*

Regression on a manifold with a Laplace eigenbasis and topological penalty

Olympio Hacquard [1] & Gilles Blanchard [1] & Clément Levrard [2] & Wolfgang Polonik [3] & Krishnakumar Balasubramanian [3]

[1] *LMO, 307 rue Michel Magat, 91400 Orsay, ohacquar@universite-paris-saclay.fr, gilles.blanchard@universite-paris-saclay.fr*
[2] *LPSM, 1 place Aurélie Nemours, 75013 Paris, clement.levrard@lpsm.paris*
[3] UC Davis, 399 Crocker Ln, Davis, CA 95616, États-Unis, wpolonik@ucdavis.edu, kbala@ucdavis.edu

Résumé. On considère un problème de regression où l'on observe des données X sur une variété et des étiquettes réelles Y, et on cherche à estimer la fonction de régression $f(x) = \mathbb{E}[Y|X = x]$. On effectue la régression sur les premières fonctions propres de l'opérateur de Laplace-Beltrami. Puisque ces fonctions sont très compliquées à estimer (en particulier lorsque la variété n'est pas connue ce qui est courant en pratique), on construit un graphe sur les données et on remplace les fonctions propres de l'opérateur de Laplace-Beltrami par les vecteurs propres de la matrice Laplacienne du graphe. Généralisation naturelle de la base de Fourier à une variété générale, les bases Laplaciennes sont composées de fonctions oscillantes et sont particulièrement propices aux phénomènes de surinterprétation. On discutera deux type de pénalités topologiques construites sur l'homologie persistante des sous-ensembles de niveaux de fonctions permettant de proposer une certaines robustesse au bruit et de pallier aux problèmes de surinterprétation.

Mots-clés. Régression statistique, Laplacien de graphe, Analyse topologique de données

Abstract. We consider a regression set-up where we observe data X lying on a manifold and real labels Y and we try to predict the regression function $f(x) = \mathbb{E}[Y|X = x]$. The regression task is performed on the first eigenfunctions of the Laplace-Beltrami operator. As these functions can be very hard to estimate (in particular when the manifold is unknown which is very frequent in practice), we build a graph on the data points and use the eigenvectors of the graph Laplacian as a new regression basis. As they are a natural generalization of Fourier bases to a manifold, Laplaces eigenbases are constituted of oscillating functions and are very likely to overfit the data. We will discuss two types of topological penalties built on the persistent homology of the sublevel sets of some functions in order to provide a robustness to noise and avoid overfitting problems.

Keywords. Statistical regression, Graph Laplacian, Topological data analysis ...

1 Modèle statistique

On observe un $n-$échantillon $(X_i, Y_i)_{i=1}^n$ où les X_i vivent sur une sous-variété de $\mathbb{R}^D$ compacte sans bord $\mathcal{M}$ et où les Y_i sont des étiquettes réelles. On suppose que les données sont simulées de sorte que pour tout i,

$$Y_i = f^\star(X_i) + \varepsilon_i$$

où ε est un bruit sous-gaussien indépendant sur chaque entrée et l'on cherche à estimer la fonction $f^\star$. Pour ce faire, on construit un graphe sur les données X_i avec des poids W_{ij} dépendant de la métrique ambiante. Les deux exemples de graphe considérés dans les expériences sont des graphes aux k plus proches voisins où $W_{ij} = 1$ si x_i est parmi les k plus proches voisins de j et 0 sinon, et des graphes gaussiens $W_{ij} = \frac{1}{n}\frac{1}{t(4\pi t)^{d/2}} e^{-\frac{\left\|X_i - X_j\right\|^2}{4t}}$ où l'on a introduit un paramètre d'échelle t. En notant D la matrice diagonale des degrés

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \text{ and } D_{ij} = 0 \text{ if } i \neq j,$$

on peut introduire la matrice Laplacienne du graphe $L = D - W$. Cette matrice est symétrique définie positive et possède ainsi une base orthonormée de vecteurs propres $(\Phi_i)_{i=1}^n$. L'utilisation d'une telle base pour effectuer diverses tâches d'apprentissage statistique a été largement étudiée depuis son introduction par Belkin et Niyogi (2003). Lorsque le nombre de points n tend vers l'infini et le paramètre d'échelle t tend vers 0, la base de vecteurs propres tend vers les fonctions propres de l'opérateur de Laplace Beltrami continu (voir Trillos & al. (2020) pour un traitement récent et complet). En figure 1 figurent quelques exemples de fonctions propres d'un graph Laplacien sur le tore.

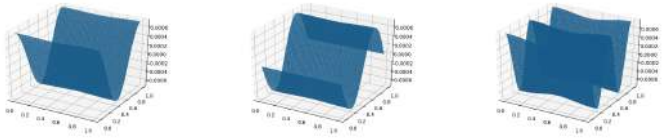
On introduit une nouvelle matrice d'apprentissage X où la $i-$ème colonne de X est le vecteur Φ_i. Le problème revient donc à trouver $\hat{\theta} \in \mathbb{R}^n$ minimisant la fonctionnelle

$$\mathcal{L}(\theta) = \|Y - X\theta\|_2^2 + \mu\Omega(\theta).$$

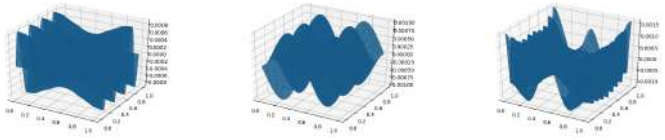
où Ω est un terme de pénalité visant à promouvoir une bonne généralisation (voir Massart (2007)).

2 Pénalités topologiques

Inspiré de travaux récents de Chen & al. (2019) et Carriere & al. (2020), on cherche à introduire un terme de régularisation topologique. Les pénalités proposées sont basées sur la notion de persistance topologique d'une fonction(voir Boissonnat, Chazal et Yvinec (2018)). Lorsque les sous-ensembles de niveau d'une fonction varient de $-\infty$ à $+\infty$, leur topologie change, et plus précisément des composantes homologiques naissent et



(a) EF 1 (b) EF 2 (c) EF 3



(d) EF 8 (e) EF 21 (f) EF 45

Figure 1: i-ème fonction propre dépliée du 20-ppv. graphe construit sur 10000 points sur le tore

meurent. La persistance est définie comme la somme des temps de vie de chaque composante topologique. Une représentation usuelle est celle des diagrammes de persistance qui est un multi-ensemble de points où pour chacun est représenté en abscisse le temps où la composante topologique est née et en ordonnée le temps où celle-ci est morte. Lorsque l'on bruite une fonction, de nombreux points sont ajoutés près de la diagonale, et le paradigme classique est de considérer que les points éloignées de la diagonale correspondent à de vraies composantes topologiques de la fonction tandis que ceux près de la diagonale correspondent à du bruit (voir figure 2). On note χ_d la persistance en dimension d et χ la persistance totale, somme des persistances en chaque dimension.

La première pénalitée considérée est :

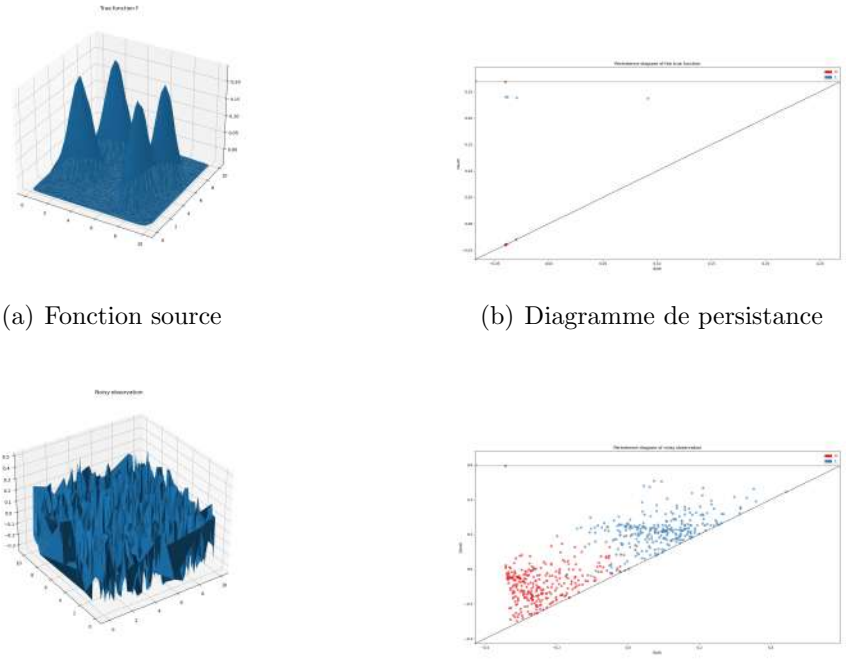
$$\Omega_1(\theta) = \sum_{i=1}^{p} |\theta_i| \chi(\Phi_i).$$

Cette pénalité est convexe en θ et pénalise chaque fonction propre individuellement, utilisant les propriétés de sélection des pénalités en norme 1 afin d'éliminer les fonctions propres qui oscillent trop et ont donc une persistance trop élevée.

La seconde pénalité est :

$$\Omega_2 = \chi \left(\sum_{i=1}^{p} \theta_i \Phi_i \right).$$

Celle-ci est non-convexe et a pour objectif de pénaliser directement la géométrie de la fonction de régression que l'on cherche à estimer, dans le but de la lisser et de réduire le bruit.



(a) Fonction source

(b) Diagramme de persistance

(c) Perturbation bruitée

(d) Diagramme de persistance

Figure 2: Influence du bruit sur le diagramme de persistance

3 Résultats expérimentaux

Les deux pénalités ont été essayées sur de nombreuses données à la fois réelles et synthétiques et ont été comparées à des pénalités plus usuelles (Lasso ou variation totale), ainsi qu'à des méthodes à noyau standard.

La pénalité Ω_1 tend à être meilleure que les pénalités usuelles sur la plupart des données synthétiques considérées et s'est montrée particulièrement efficace sur des données réelles. Un exemple consiste à considérer des images placées sur un plateau tournant et à prédire l'angle de rotation de l'objet (voir figure 3). Cette pénalité est particulièrement efficace en présence de bruit (relativement aux autres méthodes usuelles de prédiction).

La pénalité Ω_2 est également efficace en présence de bruit et permet de reconstruire des fonctions en respectant leur topologie. En figure 4 figure la reconstruction de la fonction bruitée de la figure 2. Sur cet exemple, la reconstruction est visuellement bien meilleure et plus fidèle avec une pénalité topologique : on observe bien 4 pics sur la reconstruction (et sur le diagramme de persistance) tandis qu'un Lasso ne parvient qu'à reconstruire 3 pics. De plus, la persistance de la reconstruction par Lasso est bien plus élevée que celle de la fonction reconstruite par pénalité topologique (et que la fonction source), en particulier

(a) Image de base (b) Image tournée de 30° (c) Image tournée de 90°

Figure 3: Données réelles

il y a de nombreux points proches de la diagonale dans le diagramme de persistance.

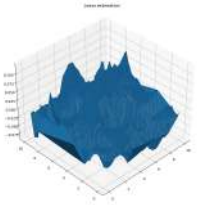
Pendant la mise en oeuvre expérimentale, on a noté qu'une méthode particulièrement efficace était de faire une sélection de variable avec la pénalité Ω_1, puis de faire tourner la régularisation Ω_2 sur les fonctions propres sélectionnées. En résumé, les pénalités Ω_1 et Ω_2 apparaissent comme complémentaires et sont particulièrement efficaces lorsque le bruit est important et le nombre de points relativement faible, où lorsque l'on cherche à généraliser à de nouvelles données. L'intérêt par rapport à une variation totale apparaît à partir de la dimension 2 puisque la topologie s'intéresse à tous les points critiques tandis que la variation totale ne permet de pénaliser que les extrema. Notons que la méthode est basée sur une décomposition spectrale du Laplacien qui utilise la structure de variété sur laquelle vivent les données, et qu'ainsi, cette méthode sera plus intéressante qu'une méthode à noyau lorsque la structure de variété est forte. Ainsi, on a également pu observer des résultats prometteurs avec des données sur un swiss roll.

4 Discussion théorique

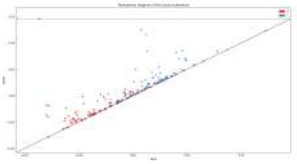
L'étude théorique de la pénalité Ω_1 est très simple : en effet, la matrice de régression X peut être choisie orthonormale, et en divisant chaque colonne par la persistance de la fonction propre correspondante, on est alors ramenés à étudier un Lasso pondéré. Toutes les propriétés connues du Lasso (voir Buhlmann & Van de Geer (2011) pour un traitement exhaustif) découlent alors. En particulier, on a

Théorème 1 *Supposons qu'il existe $\theta^0 \in \mathbb{R}^n$ tel que $f^\star(X_1, \dots, X_n) = \sum_{i=1}^n \theta_i^0 \Phi_i$, et soit $\hat{\theta}$ le minimum de la fonction $\mathcal{L}$ pour la régularisation Ω_1. Alors il existe une constante C telle que*

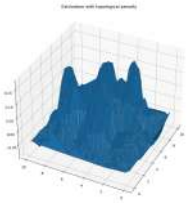
$$\|\hat{\theta} - \theta^0\|_2^2 \leq C \frac{\log n}{n}$$



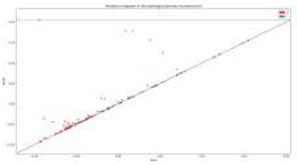
(a) Fonction estimée avec un Lasso



(b) Diagramme de persistance pour un Lasso



(c) Fonction estimée par pénalité Ω_2



(d) Diagramme de persistance estimé avec la pénalité Ω_2

Figure 4: Reconstruction topologique

De plus, la matrice de régression étant orthonormale, on dispose d'une solution explicite du Lasso pondéré :

Théorème 2 *Le minimum $\hat{\theta}$ de la fonction de perte $\mathcal{L}$ pour la régularisation Ω_1 a pour expression :*

$$(\hat{\theta}_\lambda)_j = X_j^T Y \left(1 - \frac{\lambda\chi(\Phi_j)}{2|X_j^T Y|}\right)_+.$$

En particulier, le $j$ème vecteur propre est sélectionné si et seulement si

$$|\langle X_j, Y\rangle| \leq \frac{\lambda}{2}\chi(\Phi_j).$$

A l'inverse, la pénalité Ω_2 est non-convexe et une étude poussée de cette pénalité paraît hors d'atteinte. Nous avons néanmoins pu proposer une inégalité oracle sur la prédiction :

Théorème 3 *Supposons $f^\star = \sum_{j=1}^p \theta_j^\star \Phi_j$ où les Φ_j sont les fonctions propres de l'opérateur de Laplace Beltrami pour la valeur propre λ_j. Supposons que l'on observe $Y_i = f^\star(X_i) + \varepsilon_i$*

où ε_i est un bruit sous-gaussien i.i.d. de variance v. Alors, le minimum $\hat{\theta}$ pour la pénalité Ω_2 vérifie, avec une probabilité supérieure à $1 - e^{-\kappa x}$:

$$\|\theta^\star - \hat{\theta}\|^2 \leq \frac{C_0 pv}{n}(1+\sqrt{x})^2 + C_{\mathcal{M}} C_\lambda \nu(f^\star)^2 \mu^2$$

où C_0 est une constante universelle, $C_{\mathcal{M}}$ une constante dépendant uniquement de $\mathcal{M}$ et de sa métrique, $C_\lambda = \sum_{i=1}^{p} \lambda_i^{d-1}$ et $\nu(f^\star)$ est le nombre de points dans le diagramme de persistance de $f^\star$.

Cette inégalité oracle prédit une vitesse de convergence d'ordre p/n ce qui est ce à quoi on pourrait s'attendre d'un tel modèle sans hypothèse supplémentaire de sparsité. On voit également apparaître le terme d'ordre topologique qui permet de calibrer μ. Les constantes multiplicatives dépendent de la variété elle-même, des valeurs propres du Laplacien, ainsi que de la "complexité topologique" de la fonction $f^\star$ à estimer, à savoir le nombre de points dans son diagramme de persistance.

Une question qui est survenue naturellement lors de ces travaux a été de s'intéresser au pouvoir de prédiction de classes de fonctions dont on borne la persistance. En particulier, on s'intéresse à la fat-shattering dimension de tels espaces fonctionnels. En dimension 1, on a, en utilisant un lien entre persistance et variation totale :

Théorème 4 *Soit $\mathcal{H}_V = \{f : [0,1] \to [0,1] | \chi_0(f) \leq V\}$*
Alors $fat_\gamma(\mathcal{H}_V) = 1 + \lfloor \frac{V}{4\gamma} \rfloor$

Malheureusement, en dimensions supérieures, l'ensemble des fonctions à persistence bornée est trop gros :

Théorème 5 *Soit*

$$\mathcal{H}_V = \{f : [0,1]^d \to [0,1] | \chi_0(f) \leq V_0, \chi_1(f) \leq V_1, \ldots, \chi_{d-1}(f) \leq V_{d-1}\}$$

Alors $fat_\gamma(\mathcal{H}_V) = \infty$ si $2\gamma < V_0$ et 0 sinon.

La recherche d'un espace fonctionnel $\mathcal{F}$ suffisamment régulier dont la fat-dimension (ou l'entropie métrique) serait réduite en intersectant avec $\mathcal{H}_V$ est toujours une question ouverte et serait un premier pas très intéressant dans la compréhension des pénalités topologiques, sur lesquelles aucun résultat théorique significatif n'a été énoncé jusqu'à présent.

Bibliographie

Mikhail Belkin and Partha Niyogi (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, pp.1373–1396
Nicolas Garcia Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepcev (2020) Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*,pp.827–887
Pascal Massart (2007) Concentration inequalities and model selection.
Mathieu Carriere, Frederic Chazal, Marc Glisse, Yuichi Ike, and Hariprasad Kannan (2020). A note on stochastic subgradient descent for persistence-based functionals: convergence and practical aspects *Arxiv preprint*.
Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology (2019). *22nd International Conference on Artificial Intelligence and Statistics* pp.2573–2582.
Jean-Daniel Boissonnat, Frederic Chazal, and Mariette Yvine (2018). Geometric and topological inference. *Cambridge University Press*
Peter Buhlmann and Sara Van De Geer (2011). Statistics for high-dimensional data: methods, theory and applications. *Springer Science & Business Media*.

COMPARAISON DE MODÈLES EN DÉCONVOLUTION : ÉVIDENCES, CHIB ET GIBBS

Benjamin Harroué[1,2], Jean-François Giovannelli[1] et Marcelo Pereyra[2]

[1] *Laboratoire IMS (Univ. Bordeaux – CNRS – BINP), Talence, France*

[2] *MACS, Heriot-Watt University, Edinburgh, United Kingdom*

Réesumé — Le problème de la restauration d'image présente une difficulté récurrente liée à son caractère mal-posé, requérant la prise en compte d'information complémentaire à celle apportée par les données. Dans un cadre bayésien, en particulier celui du simple filtrage de Wiener, cette information est décrite par des modèles probabilistes gaussiens et stationnaires caractérisés par leur densité spectrale de puissance. La structure de ces modèles est souvent choisie au sein d'un catalogue de manière empirique. L'intérêt pour la sélection automatique apparaît alors clairement mais elle est pourtant peu abordée dans ce cadre de l'inversion. Nous nous appuyons sur la théorie bayésienne de la décision optimale : sélection du modèle de plus forte probabilité *a posteriori* parmi les modèles du catalogue. Ces probabilités reposent sur les évidences (vraisemblances marginales) et nous abordons leur calcul par l'approche de Chib et un algorithme de Gibbs. Nous détaillons les diverses étapes du calcul ainsi que de l'algorithme et nous montrons des résultats quantitatifs probants.

Mots-clés — Sélection de modèle, stratégie bayésienne, évidence, approche de Chib, échantillonneur de Gibbs, problème inverse, déconvolution, filtrage de Wiener.

Abstract — The problem of image restoration presents a reoccurring difficulty linked to its badly-scaled character, requiring additional information to that provided by the data to be taken into account. In a Bayesian framework, in particular that of simple Wiener filtering, this information is described by Gaussian and stationary probabilistic models characterised by their power spectral density. The structure of these models is often empirically chosen from a catalog. The interest in automatic selection is then clear, but it is however little discussed in the context of inversion. We rely on the Bayesian theory of optimal decision : selection of the model with the highest posterior probability among the models in the catalog. Each probability is based on an evidence (marginal likelihood) and we address its computation by the Chib approach and a Gibbs algorithm. We detail the various steps of the calculations as well as the algorithm and show convincing quantitative results.

Keywords — Models selection, Bayesian strategy, evidence, Chib approach, Gibbs sampler, inverse problem, deconvolution, Wiener filtering.

1 Déconvolution d'image et introduction du problème

La restauration d'image est un sujet d'intérêt dans de nombreux domaines (*e.g.,* astronomie, médecine, surveillance et contrôle,...) et pour diverses modalités (*e.g.,* scanner et rayons X, optique éventuellement par interférométrie,...). Une des difficultés récurrentes provient du caractère mal-posé, requérant la prise en compte d'information en plus de celle fournie par les données. Dans un cadre bayésien, cette information est décrite par des modèles probabilistes (*e.g.,* Gauss, Poisson,... blanc ou corrélé, à structure markovienne ou plus générale,...). Dans le contexte simple du filtrage de Wiener, elle est encodée au travers de modèles gaussiens stationnaires caractérisés pas leur densité spectrale de puissance pour l'erreur et pour l'image ainsi que quelques hyperparamètres (*e.g.,* niveaux d'erreur et de signal). La structure de ces modèles est souvent choisie au sein d'un catalogue de manière empirique par essais et erreurs. Cette approche possède deux limitations : elle présente une part d'arbitraire et elle requiert une énergie importante pour réaliser les études, ce qui devient irréaliste pour de gros volumes de données. L'intérêt pour la sélection automatique apparaît alors de façon évidente mais elle est pourtant peu abordé dans le cadre de l'inversion. D'une manière générale, il existe de nombreux critères [1, 2] et nous nous appuierons sur la théorie bayésienne de la décision optimale : sélection du modèle de plus forte probabilité *a posteriori* parmi les modèles du catalogue [3]. Ces travaux résultent de [4] et sont en partie décrits dans [5, 6] (voir aussi [7]).

2 Notation et position du problème

Considérons la question de l'estimation d'une image $\boldsymbol{x} \in \mathbb{R}^P$ à partir de l'observation d'une image $\boldsymbol{y} \in \mathbb{R}^P$ liée à $\boldsymbol{x}$ par $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}$, où $\boldsymbol{H} \in \mathbb{R}^{P\times P}$ et $\boldsymbol{e} \in \mathbb{R}^P$ modélisent un flou et une erreur. On considère divers modèles pour $\boldsymbol{H}$, $\boldsymbol{x}$ et $\boldsymbol{e}$ et on s'intéresse à la comparaison quantitative de ces modèles, sur la seule base de $\boldsymbol{y}$ sans connaissance de vérité terrain.

— Plus précisément, $\boldsymbol{H}$ est une matrice de convolution associée à diverses réponses impulsionnelles, par exemple des réponses uniformes à support carré ou circulaire de tailles variées ainsi qu'un Kronecker (pas de convolution).

— $\boldsymbol{x}$ et $\boldsymbol{e}$, sont des vecteurs gaussiens de précision $\boldsymbol{P}_\star$. On s'intéresse spécifiquement au cas markovien stationnaire dont l'énergie de Gibbs est encodée par divers filtres. Les précisions associées s'écrivent alors $\boldsymbol{P}_\star = \gamma_\star \boldsymbol{C}_\star^{\mathrm{t}} \boldsymbol{C}_\star$ où $\boldsymbol{C}_\star$ est la matrice de convolution associée au filtre considéré et $\gamma_\star$ est un paramètre d'échelle.

Ces modèles capturent une large variété de situations à la fois en terme de structure de régularité ou de corrélation inter-pixels ainsi que de systèmes d'observation. On note M le nombre total de modèles constitué ainsi, par un triplet : réponse impulsionnelle, précision pour l'objet et précision pour l'erreur.

3 Probabilité, évidence, Chib et Gibbs

En substance, pour les comparer, on calcule la probabilité pour chacun des M modèles $\mathcal{M}$

$$p(\mathcal{M}=m\,|\,\boldsymbol{y}) = \frac{p(\boldsymbol{y}\,|\,\mathcal{M}=m)\,p(\mathcal{M}=m)}{p(\boldsymbol{y})}\,,$$

pour $m=1,\dots M$. La clé est la vraisemblance marginale, appelée *évidence*

$$p(\boldsymbol{y}\,|\,\mathcal{M}=m) = \iint_{\boldsymbol{\gamma},\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}\,|\,\mathcal{M}=m)\,\mathrm{d}\boldsymbol{\gamma}\,\mathrm{d}\boldsymbol{x}\,,$$

dont le calcul est délicat. Dans le cas présent, l'objet s'intègre explicitement mais pas les hyperparamètres et nous nous appuyons alors sur l'idée de Chib [8] qui repose sur le fait que :

$$p(\boldsymbol{y}\,|\,\mathcal{M}) = \frac{p(\boldsymbol{y},\boldsymbol{\gamma}\,|\,\mathcal{M})}{p(\boldsymbol{\gamma}\,|\,\boldsymbol{y},\mathcal{M})} = \frac{p(\boldsymbol{y}\,|\,\boldsymbol{\gamma},\mathcal{M})\,p(\boldsymbol{\gamma}\,|\,\mathcal{M})}{p(\boldsymbol{\gamma}\,|\,\boldsymbol{y},\mathcal{M})}\,,$$

pour tout $\boldsymbol{\gamma}$. La difficulté réside alors au dénominateur, nécessitant aussi une marginalisation mais qui se résout en l'écrivant comme une espérance approchée par une moyenne empirique

$$p(\boldsymbol{\gamma}\,|\,\boldsymbol{y},\mathcal{M}) = \int_{\boldsymbol{x}} p(\boldsymbol{\gamma},\boldsymbol{x}\,|\,\boldsymbol{y},\mathcal{M})\,\mathrm{d}\boldsymbol{x} = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\mathcal{M}}\left[p(\boldsymbol{\gamma}\,|\,\boldsymbol{x},\boldsymbol{y},\mathcal{M})\right] \simeq \frac{1}{N}\sum_{n} p(\boldsymbol{\gamma}\,|\,\boldsymbol{x}^{[n]},\boldsymbol{y},\mathcal{M})$$

ou les $\boldsymbol{x}^{[n]}$ sont des tirages de $p(\boldsymbol{x}\,|\,\boldsymbol{y},\mathcal{M})$ obtenus comme sous-produit d'un échantillonneur de Gibbs [3] pour $p(\boldsymbol{x},\boldsymbol{\gamma}\,|\,\boldsymbol{y},\mathcal{M})$ déjà donné dans [9].

Dans la présentation, nous détaillons les étapes de ces calculs ainsi que chacune des densités en jeu à partir des caractéristiques de la réponse $\boldsymbol{H}$ et des précisions $\boldsymbol{P}_{\boldsymbol{x}}$ et $\boldsymbol{P}_{\boldsymbol{e}}$. Nous rappelons également les étapes de l'algorithme de Gibbs et les différentes conditionnelles donné dans [9] (voir aussi [10, Sec. 4]). L'accent est mis sur le cas stationnaire-circulant et le contexte du filtrage de Wiener. On montre comment les calculs se mettent en œuvre entièrement dans la plan de Fourier de manière efficace en parallèle. On pourra également consulter [6].

4 Résultats

Nous présentons une première étude numérique sur données synthétiques pour évaluer la sélection proposée dans un cadre déconvolution-débruitage par filtrage de Wiener.

Nous considérons $M=32$ modèles. En ce qui concerne $\boldsymbol{H}$, nous avons 8 filtres comme indiqué sur la Fig. 1 avec une taille de 3 et 5. Pour $\boldsymbol{e}$, nous avons 4 modèles alternatifs : blanc, haute fréquence et deux modèles résonnants (voir la Figure 2). Nous considérons un seul modèle pour $\boldsymbol{x}$: une image large bande à basse fréquence (la deuxième de la Figure 2). Les valeurs

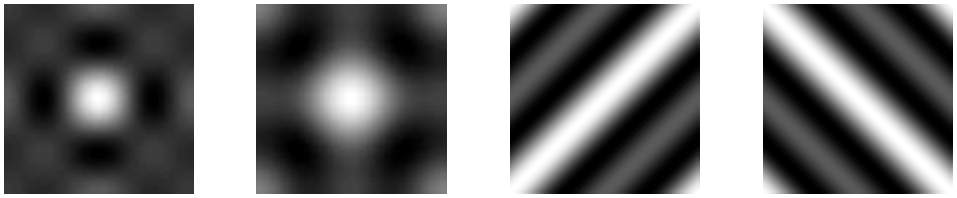


FIGURE 1 – Structure des filtres (plan de Fourier). De gauche à droite : carré, circulaire et deux mouvements diagonaux de taille 3 (une taille 5 est également utilisée).

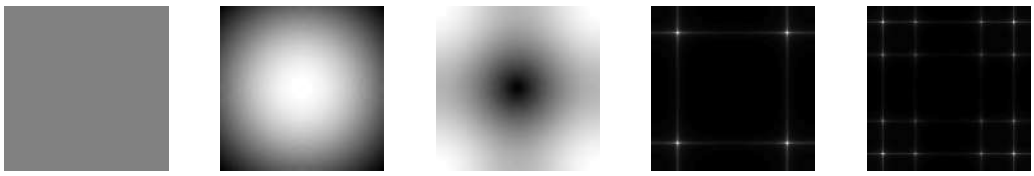


FIGURE 2 – Densités spectrales de puissance. De gauche à droite : bruit blanc, basse fréquence, haute fréquence ainsi que deux structures résonnantes.

vraies des hyperparamètres sont $\gamma_x^\star = 1$ et $\gamma_e^\star = 5$. Les images sont de taille 512×512. Pour chacun des $M = 32$ modèles, nous avons généré 100 images synthétiques floues et bruitées $\boldsymbol{y}$.

Ensuite, pour chacun des M (vrais) modèles $m^\star$ et chacune des 100 images $\boldsymbol{y}$, nous avons calculé les M probabilités $p(\mathcal{M} = m \,|\, \boldsymbol{y})$ pour $m = 1, \ldots, M$ comme décrit précédemment. Nous avons alors sélectionné le modèle le plus probable a posteriori $\widehat{m} = \arg\max_m p(\mathcal{M} = m \,|\, \boldsymbol{y})$. La Figure 3 donne les résultats sous forme de matrice de confusion.

Nous observons que les résultats sont très précis pour tous les modèles considérés, avec une précision allant de 90% à 100% selon la configuration spécifique. Les modèles 9 et 13 semblent être légèrement plus difficiles, avec des taux respectifs de 92% et 95%. Plus précisément, le modèle $m^\star = 9$ est sélectionné comme $\widehat{m} = 10$ dans 8% des instances et le modèle $m^\star = 13$ est sélectionné comme $\widehat{m} = 14$ dans 5% des instances. Les modèles 9 et 13 impliquent tous deux une erreur blanche tandis que les modèles 10 et 14 impliquent tous deux une erreur à haute fréquence à large bande (voir première et troisième de la Fig.2). Cette confusion s'explique probablement par le fait que les deux modèles d'erreur sont semblables dans les hautes fréquences et que les basses fréquences sont essentiellement occupées par une contribution venant de l'entrée.

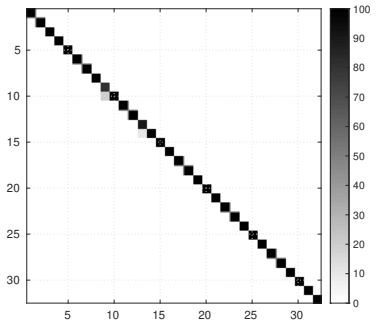


FIGURE 3 – Matrice de confusion, *i.e.,* comparaison de $m^\star$ et de $\widehat{m}$ (en pourcentage). Modèle vrai $m^\star$ (x-axis) et modèle candidat $\widehat{m}$ (y-axis).

5 Conclusion

Nous présentons une contribution originale à la sélection automatique de modèles dans un contexte déconvolution-débruitage d'images. Notre stratégie est optimale au sens d'un risque bayésien et repose sur le choix du modèle le plus probable. Ainsi, pour chaque modèle, nous évaluons la probabilité, déduite de l'évidence, résultant de la marginalisation de l'image inconnue et des hyperparamètres. Nous travaillons dans le cas gaussien permettant la marginalisation de l'image et la difficulté majeure concerne la marginalisation des hyperparamètres. Plusieurs options sont disponibles et nous nous appuyons sur l'approche de Chib et un échantillonneur de Gibbs ce qui assure la convergence. De plus, nous travaillons avec des structures circulantes ce qui autorise calculs particulièrement rapides. Nous montrons d'excellentes performances : très forte probabilité de décision correcte.

Dans une version étendue de ce travail, nous proposerons une comparaison avec d'autres méthodes existantes [1–3] : approximation de Laplace, RJMCMC, WBIC pour calculer les probabilités des modèles et les critères d'information tels que AIC ou BIC.

Parmi les perspectives, dans le cas gaussien non-circulant on aura recours à [11–14] pour l'échantillonnage des images et pour les cas non-gaussiens on s'appuiera sur des échantillonneurs plus avancés [15, 16], le reste de l'algorithme restant largement inchangé. Néanmoins, nous serons confrontés à une nouvelle difficulté liée aux hyperparamètres et aux fonctions de partition des champs. Nous avons également l'intention d'inclure de nouveaux hyperparamètres, tels que des paramètres de forme des structures de covariance de l'erreur et de l'image ainsi que de la réponse impulsionnelle. Naturellement, le traitement de données réelles fait également partie de nos projets.

Références

[1] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques : An overview," *IEEE Signal Proc. Mag.*, vol. 35, pp. 16–34, nov. 2018.

[2] T. Ando, *Bayesian model selection and statistical modeling*. Boca Raton, USA : Chapman & Hall/CRC, 2010.

[3] C. P. Robert, *The Bayesian Choice. From decision-theoretic foundations to computational implementation*. Springer Texts in Statistics, New York, USA : Springer Verlag, 2007.

[4] B. Harroué, *Approche bayésienne pour la sélection de modèle à partir de d'observations indirectes*. Thèse de Doctorat, Université de Bordeaux, Bordeaux, France, juin 2020.

[5] B. Harroué, J.-F. Giovannelli, and M. Pereyra, "Sélection de modèles en restauration d'image. Approche bayésienne dans le cas gaussien," in *Actes 27*$^{\text{e}}$ *coll. GRETSI*, (Lille, France), août 2019.

[6] B. Harroué, J.-F. Giovannelli, and M. Pereyra, "Bayesian model selection for unsupervised image deconvolution with structured Gaussian priors," in *Proc. of the Int. Conf. on Stat. Signal Proc.*, (Rio de Janeiro, Brasil), juil. 2021.

[7] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu, "Bayesian texture classification from indirect observations using fast sampling," *IEEE Trans. Signal Processing*, vol. 64, no. 1, pp. 146–159, 2016.

[8] B. P. Carlin and S. Chib, "Bayesian model choice via Markov Chain Monte Carlo methods," *J. R. Statist. Soc. B*, vol. 57, pp. 473–484, 1995.

[9] F. Orieux, J.-F. Giovannelli, and T. Rodet, "Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution," *J. Opt. Soc. Amer.*, vol. 27, pp. 1593–1607, juil. 2010.

[10] C. Vacar and J.-F. Giovannelli, "Unsupervised joint deconvolution and segmentation method for textured images : a Bayesian approach and an advanced sampling algorithm," *EURASIP Journal on Advances in Signal Processing*, jan 2019.

[11] M. Vono, N. Dobigeon, and P. Chainais, "High-dimensional Gaussian sampling : a review and a unifying approach based on a stochastic proximal point algorithm," *arXiv (2010.01510)*, oct. 2020.

[12] F. Orieux, O. Féron, and J.-F. Giovannelli, "Sampling high-dimensional Gaussian fields for general linear inverse problem," *IEEE Signal Proc. Let.*, vol. 19, pp. 251–254, mai 2012.

[13] C. Gilavert, S. Moussaoui, and J. Idier, "Efficient Gaussian sampling for solving large-scale inverse problems using MCMC," *IEEE Trans. Signal Processing*, vol. 63, pp. 70–80, jan. 2015.

[14] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet, "An auxiliary variable method for MCMC algorithms in high dimension," *Entropy*, vol. 20, p. 110, 2018.

[15] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Stat. Comput.*, mai 2015.

[16] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE J. of Selec. Topics in Signal Proc.*, vol. 20, no. 2, pp. 2385–2397, 2016.

Mesures d'importance relative par décomposition de la performance de modèles de régression

Marouane Il Idrissi [1,2,3], Bertrand Iooss [1,2,3], Vincent Chabridon [1,3]

[1] *EDF R&D, 6 Quai Watier, 78400 Chatou, France; marouane.il-idrissi@edf.fr*
[2] *Institut de Mathématiques de Toulouse, 31062, Toulouse, France*
[3] *SINCLAIR AI Laboratory, Saclay, France*

Résumé. En apprentissage statistique supervisé, les mesures d'importance relative ont pour but de quantifier de manière interprétable l'importance des covariables sur la sortie du modèle d'apprentissage, notamment en présence de dépendance entre ces covariables. Dans ce papier, deux mesures particulières (les valeurs de Shapley et les valeurs proportionnelles) sont étudiées. Ces mesures sont inspirées de deux solutions d'allocations issues de la théorie des jeux. Leurs liens avec d'autres mesures connues en régression linéaire (LMG et PMVD) sont présentés. Après une première illustration de leur formulation analytique dans le cas linéaire gaussien à deux variables, leur estimation pratique, dans un contexte de régression logistique, sur un jeu de données public de prévision des feux de fôret (Algerian Forest Fires) est proposée et discutée.

Mots-clés. Mesures d'importance, régression linéaire, régression logistique, Shapley.

Abstract. In the context of supervised statistical learning, the goal of relative importance measures is to quantify, in an interpretable manner, the importance of each input in the model output, even in the context of input dependency. In the present paper, two particular measures (Shapley values and Proportional values) are studied. These measures arise from conceptual games and allocation strategies in game theory. Here, their links with usual importance measures for linear regression (LMG and PMVD) are presented. After a first illustrative analytical derivation in a two-dimensional linear Gaussian case, their practical estimation using a logistic regression model applied to a public dataset (Algerian Forest Fires) is proposed and discussed.

Keywords. Importance measures, linear regression, logistic regression, Shapley.

1 Introduction

En apprentissage statistique, la quantification de l'importance relative vise à produire des méthodes permettant d'identifier et de mesurer l'importance des covariables par le biais de mesures interprétables. Nous nous attachons ici au modèle linéaire $Y = \beta_0 + X^\top \beta$, où $Y \in \mathbb{R}$, $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ est le vecteur aléatoire des covariables du modèle, $\beta_0 \in \mathbb{R}$ et $\beta \in \mathbb{R}^d$. On note $\mathbb{V}(X_i) = \sigma_i^2$ et $\mathrm{Cov}(X_i, X_j) = \sigma_i \sigma_j \rho_{ij}$ pour $i \in D$, $D = \{1, \ldots, d\}$. La *décomposition de la variance* de Y fournit la métrique CVD (*covariance decomposition*), qui s'écrit $\mathrm{CVD}_i = \beta_i \sigma_i \sum_{j=1}^{d} \beta_j \sigma_j \rho_{ij}$, et qui permet d'assigner une valeur d'importance

à chaque X_i, $i \in D$ (Feldman 2005). Si les covariables sont indépendantes, la part de variance de Y expliquée par chacune d'entre elles est donnée par $\beta_i^2\sigma_i^2/\mathbb{V}(Y)$. Cependant, dans le cas de covariables corrélées, cette mesure CVD peut être négative ce qui rend son interprétation, en tant que part de variance, sujette à caution.

La notion d'allocation de jeux coopératifs statistiques (Feldman 2005) permet de relier les domaines de la théorie des jeux et de l'apprentissage statistique, afin de construire des mesures d'importance interprétables. Pour un modèle total paramétrique emboîtable $\Theta(X, \beta)$ (construit avec toutes les covariables), une mesure de performance positive et faiblement monotone μ_Θ peut lui être associé et être évaluée pour chaque sous-modèle restreint aux covariables d'indices $S \subset D$, et de performance $\mu_\Theta(S)$. Une famille d'allocations pour le jeu coopératif statistique (D, μ_Θ), dites à *ordres aléatoires*, peut être définie pour chaque covariable $i \in D$ (Weber 1988):

$$\phi_i = \mathbb{E}_p\Big[\mu_\Theta\big(D \setminus S_{i-1}^r\big) - \mu_\Theta\big(D \setminus S_i^r\big)\Big] = \sum_{r \in \mathcal{R}(D)} p(r)\Big(\mu_\Theta\big(D \setminus S_{r(i)-1}^r\big) - \mu_\Theta\big(D \setminus S_{r(i)}^r\big)\Big) \quad (1)$$

où p désigne une fonction de masse de probabilité définie sur $\mathcal{R}(D)$ (l'ensemble des permutations de D), $S_i^r = \{r_j\}_{j=1}^i$ dénote l'ensemble des i$^{\text{èmes}}$ premières composantes de r ($r = (r_1, \ldots, r_d) \in \mathcal{R}(D)$), et $r(i)$ est la position de l'indice i dans r. Cette famille d'allocations permet de redistribuer la performance du modèle total à chacune de ses covariables, facilitant leur interprétation.

Dans le cadre du modèle linéaire, quatre critères permettent de définir une mesure d'importance relative admissible (Johnson and Lebreton 2004; Feldman 2005; Grömping 2007) : la *positivité* ($\forall i \in D$, $\phi_i \geq 0$), l'*exclusion* (soit $\Theta(X, \beta)$, si $\beta_i = 0$, alors $\phi_i = 0$), l'*inclusion* (soit $\Theta(X, \beta)$, si $\beta_i \neq 0$, alors $\phi_i > 0$), la *contribution totale* ($\sum_{i=1}^n \phi_i = \mu_\Theta(D)$). Pour un choix spécifique de p, les allocations définies en Eq. (1) permettent de produire des allocations candidates à mesurer l'importance relative, sous couvert du respect de ces critères.

Notre objectif est d'étudier deux cas particuliers d'allocations, inspirés de résultats généraux en théorie des jeux : les valeurs de Shapley et les valeurs proportionnelles. La Section 2 vise à définir ces mesures d'importance relative dans le cadre de jeux coopératifs statistiques et d'étudier leur admissibilité, ainsi que d'illustrer leur comportement analytique dans le cas linéaire gaussien. La Section 3 est dédiée à un cas d'utilisation sur le jeu de données Algerian Forest Fires, dans un but de classification binaire par modèle de régression logistique.

2 Valeurs de Shapley et valeurs proportionnelles

Les valeurs de Shapley (Shapley 1951) constituent une solution d'allocation pour jeux coopératifs. Elles sont l'unique solution d'une définition axiomatique garantissant le respect de quatre propriétés désirées en théorie des jeux. Dans le contexte des allocations

par modèle à ordres aléatoires (Eq. (1)), les valeurs de Shapley du jeu coopératif statistique (D, μ_Θ) sont équivalentes à choisir p comme étant uniforme sur $\mathcal{R}(D)$ (i.e., chaque ordre est équi-vraisemblable). Ainsi, $\forall r \in \mathcal{R}(D)$, $p(r) = 1/d!$, ce qui amène à la mesure d'importance relative Sh_i, $\forall i \in D$ (Eq. (1) avec $p(r) = 1/d!$). Cependant, Feldman (2005) montre que cette mesure d'importance relative particulière n'est pas admissible au sens des critères énoncés plus haut. En effet, elle ne respecte pas le critère d'*exclusion* : une covariable qui n'est pas présente dans le modèle total (i.e., son paramètre est égal à zéro) peut recevoir une part de performance si elle est corrélée avec une ou plusieurs covariables qui sont présentes dans le modèle (i.e., dont les paramètres sont différents de zéro). Cet effet a été mis en évidence en analyse de sensibilité (Iooss and Prieur 2019).

De plus, le choix d'un a priori uniforme peut être remis en question en remarquant que μ_Θ peut contenir de l'information sur l'importance relative. En effet, si pour une permutation $r = (r_1, \ldots, r_d)$, les contributions séquentielles $M_i(r) = \mu_\Theta\big(D \setminus S^r_{i-1}\big) - \mu_\Theta\big(D \setminus S^r_i\big)$ sont croissantes (i.e., $M_1(r) < M_2(r) < \cdots < M_d(r)$), alors il est probable que les éléments de r soient rangés par ordre croissant d'importance relative. En étendant les *valeurs proportionnelles* (issues des jeux coopératifs, voir Ortmann (2000)) aux jeux coopératifs statistiques, Feldman (2005) introduit la mesure PMD (*proportional marginal decomposition*) par le biais d'une autre définition de p :

$$p(r) = \frac{L(r)}{\sum_{m \in \mathcal{R}(D)} L(m)}, \quad \text{avec} \quad L(r) = \left(\prod_{S \in \{S^r_i\}_{i=1}^d} \big(\mu_\Theta(D) - \mu_\Theta(D \setminus S)\big) \right)^{-1}.$$

L'allocation sur (D, μ_Θ) ainsi définie est donnée par :

$$\mathrm{PMD}_i = \left(\sum_{m \in \mathcal{R}(D)} L(m) \right)^{-1} \sum_{r \in \mathcal{R}(D)} L(r) \Big(\mu_\Theta\big(D \setminus S^r_{r(i)-1}\big) - \mu_\Theta\big(D \setminus S^r_{r(i)}\big) \Big).$$

L'existence et l'unicité de cette fonction de masse sont garanties par une définition axiomatique qui stipule notamment que si $\mu_\Theta(D \setminus \{i\}) = \mu_\Theta(D)$ (i.e., la covariable X_i n'a aucun effet sur la performance du modèle total), alors $\mathrm{PMD}_i = 0$. La valeur de l'allocation est définie sur le jeu coopératif statistique $(D \setminus \{i\}, \mu_\Theta)$, et X_i reçoit une part nulle. Ceci permet de garantir qu'une covariable non-présente dans le modèle total ne reçoive aucune contribution, malgré le fait qu'elle puisse être corrélée aux variables présentes. Cette mesure d'importance relative est *admissible* (cf. Section 1).

Appliquées au modèle de régression linéaire, avec le coefficient de détermination R^2 comme mesure de performance, les valeurs de Shapley du jeu coopératif statistique (D, R^2) sont connus comme étant les indices LMG (Lindeman, Merenda, and Gold 1980). Les valeurs proportionnelles de (D, R^2), quant à elles, sont connues sous le nom de PMVD (*proportional marginal variance decomposition*).

Pour un modèle linéaire à deux covariables $Y = \beta_1 X_1 + \beta_2 X_2$, avec $(X_1, X_2)^\top$ un vecteur gaussien centré vérifiant pour $i = 1, 2$, $\mathbb{V}(X_i) = \sigma_i^2$ et $\mathrm{Cov}(X_1, X_2) = \sigma_1 \sigma_2 \rho$,

associé au coefficient de détermination R^2 comme mesure de performance, les mesures LMG et PMVD sont données analytiquement par :

$$\begin{array}{rclcrcl} \mathrm{LMG}_1 & = & \frac{1}{\mathbb{V}(Y)}\left(\beta_1^2\sigma_1^2 + \beta_1\beta_2\sigma_1\sigma_2\rho + \frac{\rho^2}{2}(\beta_2^2\sigma_2^2 - \beta_1^2\sigma_1^2)\right) ; & \mathrm{PMVD}_1 & = & \frac{\beta_1^2\sigma_1^2}{\beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2}; \\ \mathrm{LMG}_2 & = & \frac{1}{\mathbb{V}(Y)}\left(\beta_2^2\sigma_2^2 + \beta_1\beta_2\sigma_1\sigma_2\rho + \frac{\rho^2}{2}(\beta_1^2\sigma_1^2 - \beta_2^2\sigma_2^2)\right) ; & \mathrm{PMVD}_2 & = & \frac{\beta_2^2\sigma_2^2}{\beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2} \end{array}$$

vérifiant la relation $\mathrm{LMG}_1 + \mathrm{LMG}_2 = \mathrm{PMVD}_1 + \mathrm{PMVD}_2 = R^2(\{1,2\}) = 1$. Dans ce cas précis, les PMVD ne dépendent pas de ρ (ce comportement ne se généralise pas pour $d > 2$), alors que la mesure LMG semble partager la part de variance due à la corrélation équitablement entre X_1 et X_2. De plus, si $\beta_1^2\sigma_1^2 = \beta_2^2\sigma_2^2$, les quatres indices associés aux covariables se confondent. Dans le cas où $\rho = 0$, les deux mesures se comportent de la même façon. Lorsque l'un des deux paramètres $\beta_i = 0$, alors $\mathrm{PMVD}_i = 0$ tandis que LMG_i peut être non-nul, pour des valeurs de ρ non-nulles : ce comportement est contraire au critère d'exclusion.

Dans les cas à plus de deux covariables, l'étude analytique de ces mesures proposée par Grömping (2007) permet de conclure que la mesure LMG aura tendance à répartir les effets dus à la corrélation équitablement entre les covariables, indépendamment de leur importance dans le modèle, tandis que la mesure PMVD aura tendance à attribuer ces effets aux covariables les plus importantes.

3 Prévision des feux de forêt

Dans cette section, nous étendons l'utilisation des mesures d'importance précédentes au cadre du modèle linéaire généralisé pour la régression logistique et les appliquons à un jeu de données public.

Le jeu de données Algerian Forest Fires (Abid and Izeboudjen 2020) contient $n = 244$ observations journalières enregistrées dans deux régions d'Algérie (Bejaia et Sidi Bel-abbes) entre les mois de juin et septembre 2012. Les 8 covariables sont Temp (température maximale en degrés Celsius), RH (humidité relative en %), Ws (vitesse du vent en km/h), Rain (pluviométrie totale en mm), FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code) et ISI (Initial Spread Index). Comme illustré en Figure 1, ces covariables peuvent être très corrélées les unes aux autres. Nous cherchons à modéliser la probabilité qu'un feu de forêt ait eu lieu par régression logistique et nous prenons comme mesure de performance le coefficient de détermination généralisé, qui dans ce cas vaut :

$$R^2(S) = 1 - \frac{\text{déviance du sous-modèle d'indices dans } S}{\text{déviance du modèle nul}}.$$

Estimé sur le jeu de données, nous obtenons $\widehat{R^2} \simeq 0.803$ et un coefficient de prédictivité Q^2 (R^2 en prédiction), calculé par validation croisée égal à $\widehat{Q^2} \simeq 0.79$. Les mesures de multicolinéarité VIF (*variance inflation factor*)[1] ont également été calculées, en plus

1. Fonction vif() du package car sous R.

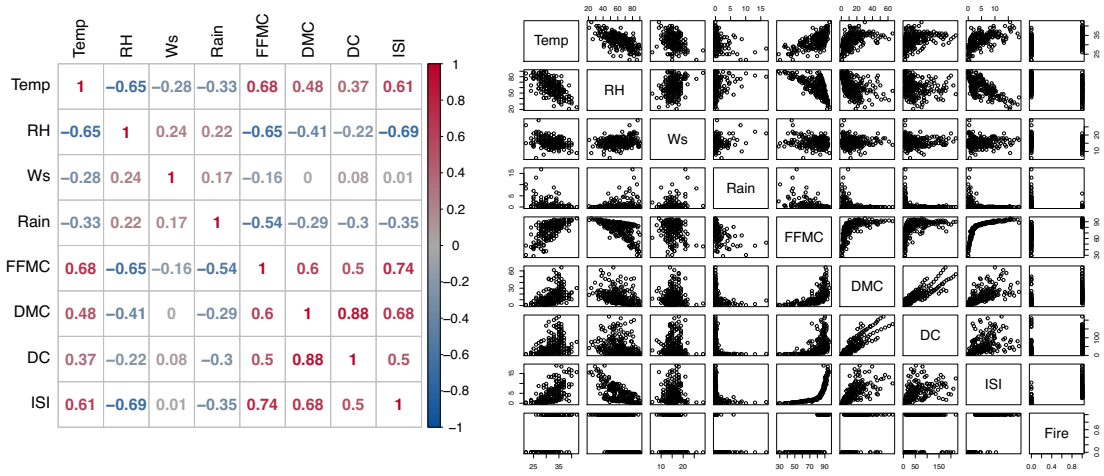


Figure 1: Matrice des corrélations (gauche) et nuages de points croisés (droite) des covariables du jeu de données Algerian Forest Fires.

Covariables	Temp	RH	Ws	Rain	FFMC	DMC	DC	ISI	Total
VIF	1.36	1.90	1.72	1.44	7.08	8.04	6.24	5.04	-
Sh (%)	4.5	3.7	0.4	5.5	33.3	6.2	3.2	23.5	80.3
PMD (%)	0.4	0	0	0.7	69.7	6.4	0	3.1	80.3

Table 1: Mesure de colinéarité et mesures d'importance relative du modèle de régression logistique sur les données Algerian Forest Fire.

des mesures d'importance par valeurs de Shapley et par valeurs proportionnelles[2]. Les résultats se trouvent en Table 1.

Les covariables FFMC, DMC, DC et ISI présentent de fortes valeurs de VIF (i.e., supérieures à 5), ce qui fait écho aux valeurs élevées des corrélations déjà identifiées en Figure 1. La mesure Sh semble favoriser les covariables FFMC et ISI, avec des part respectives de 33.3% et 23.5% de la performance totale. Les parts de performance des autres covariables oscillent entre 3.7% et 6.2%, sauf pour la vitesse du vent, avec une part inférieur. La mesure PMD, quant à elle, attribue une part de près de 69.7% de la performance à la covariable FFMC, avec des contributions à la performance de 6.4% et 3.1% respectivement aux covariables DMC et ISI. Les autres covariables reçoivent moins de 1% de performance. Ceci peut-être expliqué par les fortes corrélations dont les effets sur la performance sont principalement attribués aux covariables ayant un paramètre estimé élevé (en valeur absolue). La mesure PMD permet donc de détecter les covari-

2. Fonctions lmg() et emvd() du package sensitivity sous R.

ables les plus importantes de manière plus prononcée que la mesure Sh qui aura tendance à lisser l'importance par répartition des effets de corrélation. La covariable RH en est un exemple parlant : étant fortement corrélée linéairement avec FFMC (-0.65) et ISI (-0.69), les valeurs de Shapley auront tendance à lui accorder de l'importance (3.7% de la performance), alors que les valeurs proportionnelles indiquent que son importance est nulle.

En conclusion, nous pouvons donc retenir que, dans le contexte de la régression logistique (resp. linéaire), et en choisissant le R^2 comme critère de performance, les mesures d'importance PMD (resp. PMVD) présentent l'intérêt de remplir les quatre critères d'admissibilité d'une mesure d'importance interprétable contrairement à ses homologues que sont les valeurs de Shapley (resp. LMG).

Nous remercions Nicolas Bousquet (EDF R&D) grâce à qui ce travail a pu être réalisé.

Bibliographie

Abid, F., and N. Izeboudjen. 2020. "Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm." In *Advanced Intelligent Systems for Sustainable Development,* 363–370. Springer International Publishing.

Feldman, B. 2005. "Relative Importance and Value." *SSRN Electronic Journal.*

Grömping, U. 2007. "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition." *The American Statistician* 61 (2): 139–147.

Iooss, B., and C. Prieur. 2019. "Shapley Effects For Sensitivity Analysis With Correlated Inputs: Comparisons With Sobol' Indices, Numerical Estimation And Applications." *International Journal for Uncertainty Quantification* 9 (5): 493–514.

Johnson, J. W., and J. M. Lebreton. 2004. "History and Use of Relative Importance Indices in Organizational Research." *Organizational Research Methods* 7:238–257.

Lindeman, R. H., P. F. Merenda, and R. Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis.* Scott, Foresman.

Ortmann, K. M. 2000. "The proportional value for positive cooperative games." *Mathematical Methods of Operations Research* 51 (2): 235–248.

Shapley, L. S. 1951. *Notes on the n-Person Game – II: The Value of an n-Person Game.* Research Memorandum ATI 210720. RAND Corporation.

Weber, R. J. 1988. "Probabilistic values for games." In *The Shapley Value,* 1st ed., 101–120. Cambridge University Press.

A Bayesian Fisher-EM algorithm for discriminative Gaussian subspace clustering

Nicolas Jouvin [1] & Charles Bouveyron [2] & Pierre Latouche [3]

[1] *Institut Camille Jordan - Ecole Centrale Lyon, 69130 Ecully, FRANCE*
nicolas.jouvin@ec-lyon.fr
[2] *MAP5 - Université de Paris, 75005, Paris, FRANCE*
pierre.latouche@u-paris.fr
[3] *J.A. Dieudonné - Université Côte d'Azur, Nice, 06000 FRANCE*
charles.bouveyron@univ-cotedazur.fr

Résumé. Nous proposons une extension au cadre Bayésien du modèle de mélange Gaussien discriminant pour le clustering de données en grande dimension. Modélisant les données comme un mélange de Gaussiennes dans un sous-espace discriminant de faible dimension, un a priori gaussienn est introduit sur les moyennes de l'espace latent et une famille de douze sous-modèles est dérivée en considérant différentes structures de covariance. L'inférence des paramètres est faites via un algorithme EM variationnel, tandis que le sous-espace discriminant est estimé via la maximisation d'un critère de Fisher non supervisé. L'estimation des hyper-paramètres du modèle se fait par maximum de vraisemblance de type-II et un critère de vraisemblance classifiante intégrée est proposé pour sélectionner à la fois le nombre de groupes et le sous-modèle. L'algorithme Fisher-EM Bayésien résultant est évalué dans deux scénarios de simulation, montrant sa supériorité par rapport à l'état de l'art en clustering à l'aide de mélanges gaussiens à sous-espace latent. Une implémentation de ce travail est disponible pour le logiciel **R** dans le package **FisherEM** [1].

Mots-clés. Classification non-supervisée, Réduction de la dimension, Modèles de mélange, Inférence variationelle, Analyse discriminante de Fisher ...

Abstract In this work, we extend the powerful discriminative latent mixture model for the clustering of high-dimensional data to the Bayesian framework. Modeling data as a mixture of Gaussians in a low-dimensional discriminative subspace, a Gaussian prior distribution is introduced over the latent group means and a family of twelve submodels are derived considering different covariance structures. Model inference is done with a variational EM algorithm, while the discriminative subspace is estimated via a Fisher-step maximizing an unsupervised Fisher criterion. An empirical Bayes procedure is proposed for the estimation of the prior hyper-parameters, and an integrated classification likelihood criterion is derived for selecting both the number of clusters and the submodel. The performances of the resulting Bayesian Fisher-EM algorithm are investigated in two

[1] Disponible sur CRAN, voir https://github.com/nicolasJouvin/FisherEM pour plus d'informations.

thorough simulated scenarios, assessing its superiority with respect to state-of-the-art Gaussian subspace clustering models. This work comes with a reference implementation for the **R** software in the **FisherEM** package[2].

Keywords. Clustering, Dimension reduction, Mixture model, Variational Inference, Linear discriminant analysis

1 Introduction

Clustering has become an important part of contemporary statistics and machine learning, with applications ranging from DNA microarray analysis in biology (Ghosh and Chinnaiyan 2002) to text analysis (Aggarwal and Zhai 2012) and image processing (Jégou et al. 2010). Gaussian mixture models (GMM) constitute one of the most popular approaches to model-based clustering (McLachlan and Peel 2004). Let us consider n continuous observations $\{\boldsymbol{y}_i\}$ in dimension p, summarized in a data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, that we want to cluster into K groups. In this context, each observation $\boldsymbol{y}_i$ is assigned to a discrete latent variable $\boldsymbol{z}_i$ characterizing its cluster assignment. Gaussian mixtures posits the following distribution:

$$\boldsymbol{z}_i \sim \mathcal{M}_K(1, \boldsymbol{\pi}), \qquad \boldsymbol{y}_i \mid \{z_{ik} = 1\} \sim \mathcal{N}_p(\boldsymbol{y}_i \mid \boldsymbol{m}_k, \boldsymbol{S}_k), \tag{1}$$

where $\boldsymbol{\pi}$ denotes the mixture proportions and $(\boldsymbol{m}_k, \boldsymbol{S}_k)$ respectively corresponds to the mean and covariance matrix of the k-th component. However, the latter involve a number of parameters growing with the square of the dimension, and the number of observations required to fit high-dimensional data may be very large and computationally impractical. This is sometimes referred to as a form of *curse of dimensionality* (Bouveyron et al. 2019).

On the one hand, some approaches rely on unsupervised dimension reduction such as principal component analysis to project the data prior to model fitting (Ghosh and Chinnaiyan 2002). On the other hand, a wealth of literature has focused on developing parsimonious models based on constrained covariance matrices $\boldsymbol{S}_k$. Aside from standard spectral constraints (Banfield and Raftery 1993), other types of restrictions were considered with low-rank factorizations $\boldsymbol{S}_k = \boldsymbol{U}_k\boldsymbol{U}_k^\top + \boldsymbol{\Psi}_k$ where $\boldsymbol{U}_k$ is a $p \times d$ matrix (Ghahramani and Hinton 1996; Tipping and Bishop 1999). Based on a factor analysis formulation, these models have a geometric interpretation: integrating model-based clustering and probabilistic linear dimension reduction, they seek to cluster the data in K low-dimensional subspaces $\boldsymbol{U}_k$ of dimension d. This formulation was refined to impose a common subspace $\boldsymbol{U}$ accross cluster, further restricting the number of parameter and allowing a common vizualisation of the data points (Yoshida et al. 2004; Baek et al. 2009; McNicholas and Murphy 2008; Montanari and Viroli 2010). These models are often refered to as Gaussian subspace clustering, and maximum likelihood inference is always

[2]Available on CRAN, see https://github.com/nicolasJouvin/FisherEM for additional information.

preferred, usually via an EM algorithm. We refer the reader to Bouveyron and Brunet-Saumard (2014) for a thorough review of model-based high-dimensional clustering.

However, without further clustering information, the estimated latent subspace may be biased toward density estimation, preserving the variance of the observed data as much as possible, rather than clustering and explicit separation of the groups. These objectives are not always aligned as demonstrated by Chang (1983) with a 2-component GMM where the last principal component, retaining the least of the variance, is the best discriminative in term of cluster memberships. In order to circumvent this issue, several works proposed to chose $\boldsymbol{U}$ as the best subspace discriminating the K clusters in the sense of Fisher's Linear discriminant analysis (LDA) (Torre and Kanade 2006; Scrucca 2010; Bouveyron and Brunet 2012). The most popular criterion for this task is:

$$\boldsymbol{U}^\star = \arg\max_{\boldsymbol{U}} \left\{ \mathrm{F}(\boldsymbol{U}) \coloneqq \mathrm{Tr}\left[(\boldsymbol{U}^\top \boldsymbol{S}_T \boldsymbol{U})^{-1} \boldsymbol{U}^\top \boldsymbol{S}_B \boldsymbol{U} \right] \right\}, \tag{2}$$

where $\boldsymbol{S}_T = (1/n)\sum_i (\boldsymbol{y}_i - \boldsymbol{m})(\boldsymbol{y}_i - \boldsymbol{m})^\top$ is the empirical covariance matrix of the data, while $\boldsymbol{S}_B = (1/n)\sum_k n_k (\boldsymbol{m}_k - \bar{\boldsymbol{y}})(\boldsymbol{m}_k - \bar{\boldsymbol{y}})^\top$ is the between-class covariance, with $n_k = \sum_i z_{ik}$ and $\boldsymbol{m}_k = (1/n_k)\sum_i z_{ik}\boldsymbol{y}_i$.

2 The Bayesian discriminative latent mixture

Bouveyron and Brunet (2012) proposed the discriminative latent mixture (DLM), which consists in a standard low-rank decomposition with the additional assumption that the common subspace $\boldsymbol{U}$ is discriminative in the sense of Equation (2). The model is equivalent to Equation (1) with $\boldsymbol{m}_k = \boldsymbol{U}\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{U} + \boldsymbol{\Psi}_k$ where $\boldsymbol{U}$ is column orthornormal:

$$\boldsymbol{y}_i \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \mathcal{N}_p(\boldsymbol{U}\mu_k, \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{U} + \boldsymbol{\Psi}_k), \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d. \tag{DLM}$$

Writing down the spectral decomposition of the covariance, $\boldsymbol{S}_k = \boldsymbol{D}\boldsymbol{\Delta}_k\boldsymbol{D}^\top$, and denoting $\boldsymbol{U}^\perp$ the orthogonal complement of $\boldsymbol{U}$ in $\mathbb{R}^p$, it is further assumed that

$$\boldsymbol{\Delta}_k = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_k & 0 \\ \hline 0 & \beta_k \boldsymbol{I}_{p-d} \end{array} \right), \qquad \boldsymbol{D} = \left[\boldsymbol{U}, \boldsymbol{U}^\perp \right]. \tag{3}$$

Such decomposition amounts to consider that all the relevant clustering signal is contained in the subspace spanned by the columns of $\boldsymbol{U}$, while the orthogonal complement contains isotropic Gaussian noise with variance β_k. Then, the Fisher-EM proposed by Bouveyron and Brunet (2012) alternates between parameter estimation via maximum-likelihood and latent subspace $\boldsymbol{U}$ estimation via Equation (2), using the current posterior cluster membership probabilities to compute the scatter matrix $\boldsymbol{S}_B$, until convergence. However, while

efficient, this algorithm displays some instabilities due to poor conditioning of the scatter matrices, and can get stuck in poor local maxima in terms of clustering as demonstrated in Section 4.

In order to circumvent this issue, we propose a Bayesian extension of the DLM model where a Gaussian prior distribution is put on $\boldsymbol{\mu}_k$ as in the standard Bayesian Gaussian mixture model:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_k, \quad \boldsymbol{\mu}_k \overset{i.i.d.}{\sim} \mathcal{N}_d(\boldsymbol{\nu}, \lambda \boldsymbol{I}_d), \tag{BDLM}$$

Here, λ is a hyper-parameter controlling the spreading of the $\boldsymbol{\mu}_k$'s in the latent space and the rest of the model and assumptions is unchanged. We refer to this Bayesian version as $\text{BDLM}_{[\boldsymbol{\Sigma}_k \beta_k]}$. The set of parameters is then $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{\beta})$ and the next Section discusses inference and clustering.

A family of submodels Considering specific constraints on the matrix $\boldsymbol{\Delta}_k$, we can derive a family of submodels for the BDLM as in the original DLM. Akin to the spectral constraints of Banfield and Raftery (1993), we can assume a combination of hypotheses on the structure of the latent space covariance $\boldsymbol{\Sigma}_k$ and on the noise covariance $\boldsymbol{\Psi}_k$. First homoscedasticity constraints of the type $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ or $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$ may be considered. Moreover, the latent covariance $\boldsymbol{\Sigma}_k$ can be further assumed to be diagonal $\boldsymbol{\Sigma}_k = \text{diag}(\alpha_{k1}^2, \ldots, \alpha_{kd}^2)$ or even isotropic $\boldsymbol{\Sigma}_k = \alpha_k^2 \boldsymbol{I}_d$. The combination of these constraints leave a total of 12 submodels.

3 The Bayesian Fisher-EM algorithm

In the following, we propose a clustering algorithm based on the joint maximization of the Fisher criterion and the observed-data likelihood. Contrary to the DLM, the latter is intractable, and so is the posterior distribution of the latent variables $(\boldsymbol{Z}, \boldsymbol{\mu})$ given the data and the parameters. Thus, we propose to use variational inference, relying on a mean-field approximation of the posterior and the maximization of a lower bound of the log-likelihood (Jaakkola and Jordan 2000). The proposed clustering algorithm for BDLM is named Bayesian Fisher-EM (BFEM) and alternates between 3 steps.

Variational inference (VE-step) Introducing a *variational* distribution $q(\boldsymbol{\mu}, \boldsymbol{Z})$, the classical identity holds for any distribution q:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\vartheta}) = \underbrace{\mathbb{E}_q\left[\log p(\boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{Z} \mid \boldsymbol{\vartheta})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\mu}, \boldsymbol{Z})\right]}_{\mathcal{J}(q, \boldsymbol{\vartheta})} + \text{KL}(q \,\|\, p(\cdot \mid \boldsymbol{Y}, \boldsymbol{\vartheta})). \tag{4}$$

Here, KL denotes the Kullback-Leibler divergence, and the latter being positive, $\mathcal{J}$ is a *lower bound* of the observed data likelihood. Then, restricting q to be in the mean-field family $\mathcal{Q}$ of fully factorized distributions over $(\boldsymbol{Z}, \boldsymbol{\mu})$, we can use the standard coordinate

ascent variational inference (CAVI, Blei et al. 2017), which is a fixed point algorithm solving:

$$q^\star = \arg\max_{q \in \mathcal{Q}} \mathcal{J}(q; , \boldsymbol{\vartheta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q \,\|\, p(\cdot \mid \boldsymbol{Y}, \boldsymbol{\vartheta})). \tag{5}$$

Here, it clearly appears that the variational distribution approximates the intractable posterior through the KL minimization program. Moreover, the optimal form of the updates $q^\star$ are available in closed form in this model, as a product of Multinomials and Gaussian distributions, and can be estimated iteratively in a fixed point algorithm.

The M-step In the M-step, the bound $\mathcal{J}(\cdot, q^\star)$ is maximized with respect to the latent space mixture parameters $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Note that $\boldsymbol{U}$ is treated as a fixed, distinct parameter here. At iteration (t), the mixture proportions are estimated classically as in any other mixture models, $\hat{\pi}_k^{(t)} = \tilde{n}_k^{(t)}/n$, and the estimates of the remaining parameters $(\boldsymbol{\Sigma}_k, \beta_k)$ depend on the chosen submodel. For the unconstrained submodel, the M-step estimates of $\boldsymbol{\Sigma}_k$ and β_k are:

$$\hat{\boldsymbol{\Sigma}}_k^{(t)} = \boldsymbol{U}^\top \hat{\boldsymbol{C}}_k^{(t)} \boldsymbol{U}, \qquad \hat{\beta}_k^{(t)} = \frac{\mathrm{Tr}\left[\hat{\boldsymbol{C}}_k^{(t)}\right] - \mathrm{Tr}\left[\boldsymbol{U}^\top \hat{\boldsymbol{C}}_k^{(t)} \boldsymbol{U}\right]}{p - d}.$$

Here, $\boldsymbol{u}_h$ denotes the h-th column of $\boldsymbol{U}$ which is computed in the Fisher step (see below) at iteration (t) and:

$$\hat{\boldsymbol{C}}_k^{(t)} = \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} (\boldsymbol{y}_i - \boldsymbol{U}\tilde{\boldsymbol{\mu}}_k^{(t)})(\boldsymbol{y}_i - \boldsymbol{U}\tilde{\boldsymbol{\mu}}_k^{(t)})^\top + \boldsymbol{U}\tilde{\mathbf{M}}_k^{(t)}\boldsymbol{U}^\top,$$

where $\tau_{ik} = q^\star(z_{ik} = 1)$ denotes the variational distribution of $\boldsymbol{z}_i$.

The Fisher step (F-step) As explained above, the subspace $\boldsymbol{U}$ is supposed to be discriminative in the sense of Fisher criterion. The partition $\boldsymbol{Z}$ being unknown, the scatter matrix $\boldsymbol{S}_B$ in Equation (2) cannot be formed. Following Bouveyron and Brunet (2012) we propose to replace it by the soft between-class scatter matrix:

$$\tilde{\boldsymbol{m}}_k^{(t)} = \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{y}_i, \qquad \tilde{\boldsymbol{S}}_B^{(t+1)} = \frac{1}{n} \sum_{k=1}^{K} \tilde{n}_k^{(t)} \left(\tilde{\boldsymbol{m}}_k^{(t)} - \bar{y}\right) \left(\tilde{\boldsymbol{m}}_k^{(t)} - \bar{y}\right)^\top,$$

The unconstrained problem of Equation (2) can be cast as a generalized eigenvalue problem with an explicit solution (Ghojogh et al. 2019). Unfortunately, there is no closed-form solution for the problem with the additional orthonormality constraint $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_d$. In the supervised literature, this problem is known as orthonormal LDA and several algorithms were proposed, either iterative (Foley and Sammon 1975; Hamamoto et al. 1991) or direct (Ye 2005; Lu et al. 2016). In the line of Bouveyron and Brunet (2012), we propose to use the orthogonal discriminant vectors (ODV) method. Starting from $\boldsymbol{u}_1$, the

leading eigenvector of $\boldsymbol{S}_T^{-1}\tilde{\boldsymbol{S}}_B^{(t)}$, we greedily maximize the Fisher criterion by computing the r-th direction as the solution of the unconstrained problem in the orthogonal of the current subspace $\mathcal{B}_{r-1} = \text{vect}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r-1})$.

Note that, due to the F-step, the variational bound is not guaranteed to increase at each step. Still, we can use standard convergence criteria such as Aitken's criterion (McLachlan and Krishnan 2007, p. 147) and the BFEM alternates between the VE, M and F steps until convergence or a user-defined number of iteration is reached.

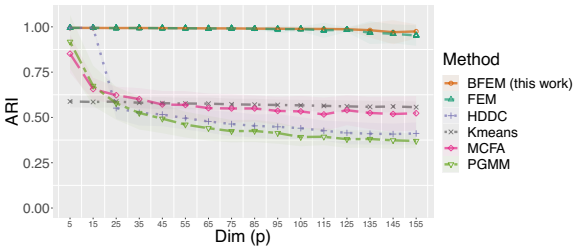
Model selection In a clustering perspective, we propose to rely on the integrated classification likelihood (ICL, Biernacki et al. 2000) to choose K and the submodel. The criterion is defined as:

$$\text{ICL}_{BIC}(\mathcal{M}, K) = \log p(\boldsymbol{Y}, \hat{\boldsymbol{Z}} \mid \hat{\boldsymbol{\vartheta}}, \mathcal{M}, K) - \frac{\gamma_{\mathcal{M},K}}{2}\log(n), \tag{6}$$

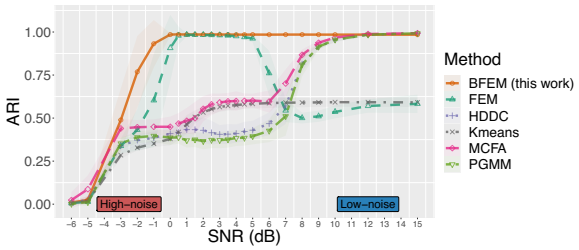
where we take $\hat{\boldsymbol{\vartheta}}$ to be the parameter estimates at the end of BFEM, and $\gamma_{\mathcal{M},K}$ is the number of free parameters in submodel $\mathcal{M}$ with K clusters. Although the marginal likelihood is intractable, the first term above is the classification likelihood which is tractable in the BDLM models. The latent dimension is fixed to $d = K - 1$ as in regular LDA.

4 Numerical results

We illustrate the performance of the proposed BFEM algorithm in two simulation scenario consisting of a $n_k = 900$ observations from a 3-components mixture in latent dimension $d = 2$ with fixed covariance matrices $\boldsymbol{\Sigma}_k$. The first scenario fixes the variance of the noise β_k and increases the dimensionality p of the problem by adding noisy dimensions. The second scenario fixes the dimension $p = 150$ and study the impact of the noise variance β_k expressed in logarithmic scale via the signal-to-noise ratio (SNR): the higher the SNR, the lower β_k. We use the *adjusted Rand index* (ARI, Hubert and Arabie 1985) in order to compare the estimated partition to the ground truth. In both scenarios, we can see that BFEM display a great stability and robustness compared to both the frequentist FEM and state-of-the-art Gaussian subspace clustering.



(a) Scenario 1: high-dimensional problems



(b) Scenario 2: robustness to noise

References

Aggarwal, Charu C and ChengXiang Zhai (2012). "A survey of text clustering algorithms". In: *Mining text data*. Springer, pp. 77–128 (cit. on p. 2).

Baek, Jangsun, Geoffrey J McLachlan, and Lloyd K Flack (2009). "Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data". In: *IEEE transactions on pattern analysis and machine intelligence* 32.7, pp. 1298–1309 (cit. on p. 2).

Banfield, Jeffrey D and Adrian E Raftery (1993). "Model-based Gaussian and non-Gaussian clustering". In: *Biometrics*, pp. 803–821 (cit. on pp. 2, 4).

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725 (cit. on p. 6).

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877 (cit. on p. 5).

Bouveyron, Charles and Camille Brunet (2012). "Simultaneous model-based clustering and visualization in the Fisher discriminative subspace". In: *Statistics and Computing* 22.1, pp. 301–324 (cit. on pp. 3, 5).

Bouveyron, Charles and Camille Brunet-Saumard (2014). "Model-based clustering of high-dimensional data: A review". In: *Computational Statistics & Data Analysis* 71, pp. 52–78 (cit. on p. 3).

Bouveyron, Charles et al. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Vol. 50. Cambridge University Press (cit. on p. 2).

Chang, Wei-Chien (1983). "On using principal components before separating a mixture of two multivariate normal distributions". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 32.3, pp. 267–275 (cit. on p. 3).

Foley, Donald H. and John W Sammon (1975). "An optimal set of discriminant vectors". In: *IEEE Transactions on computers* 100.3, pp. 281–289 (cit. on p. 5).

Ghahramani, Zoubin and Geoffrey E Hinton (1996). *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto (cit. on p. 2).

Ghojogh, Benyamin, Fakhri Karray, and Mark Crowley (2019). "Eigenvalue and generalized eigenvalue problems: Tutorial". In: *arXiv preprint arXiv:1903.11240* (cit. on p. 5).

Ghosh, Debashis and Arul M Chinnaiyan (2002). "Mixture modelling of gene expression data from microarray experiments". In: *Bioinformatics* 18.2, pp. 275–286 (cit. on p. 2).

Hamamoto, Yoshihiko et al. (1991). "A note on the orthonormal discriminant vector method for feature extraction". In: *Pattern recognition* 24.7, pp. 681–684 (cit. on p. 5).

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of classification* 2.1, pp. 193–218 (cit. on p. 6).

Jaakkola, Tommi S and Michael I Jordan (2000). "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1, pp. 25–37 (cit. on p. 4).

Jégou, Hervé et al. (2010). "Aggregating local descriptors into a compact image representation". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 3304–3311 (cit. on p. 2).

Lu, Gui-Fu, Jian Zou, and Yong Wang (2016). "A new and fast implementation of orthogonal LDA algorithm and its incremental extension". In: *Neural Processing Letters* 43.3, pp. 687–707 (cit. on p. 5).

McLachlan, Geoffrey J and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons (cit. on p. 6).

McLachlan, Geoffrey J and David Peel (2004). *Finite mixture models*. John Wiley & Sons (cit. on p. 2).

McNicholas, Paul David and Thomas Brendan Murphy (2008). "Parsimonious Gaussian mixture models". In: *Statistics and Computing* 18.3, pp. 285–296 (cit. on p. 2).

Montanari, Angela and Cinzia Viroli (2010). "Heteroscedastic factor mixture analysis". In: *Statistical Modelling* 10.4, pp. 441–460 (cit. on p. 2).

Scrucca, Luca (2010). "Dimension reduction for model-based clustering". In: *Statistics and Computing* 20.4, pp. 471–484 (cit. on p. 3).

Tipping, Michael E and Christopher M Bishop (1999). "Mixtures of probabilistic principal component analyzers". In: *Neural computation* 11.2, pp. 443–482 (cit. on p. 2).

Torre, Fernando De la and Takeo Kanade (2006). "Discriminative cluster analysis". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 241–248 (cit. on p. 3).

Ye, Jieping (2005). "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems". In: *Journal of Machine Learning Research* 6.Apr, pp. 483–502 (cit. on p. 5).

Yoshida, Ryo, Tomoyuki Higuchi, and Seiya Imoto (2004). "A mixed factors model for dimension reduction and extraction of a group structure in gene expression data". In: *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.* IEEE, pp. 161–172 (cit. on p. 2).

General Hannan and Quinn Criterion for Common Time Series

Kare KAMILA

SAMM, Université Paris 1 Panthéon-Sorbonne, FRANCE
Email: kamilakare@gmail.com

Résumé. Ce papier vise à étudier les critères de sélection de modèles dits "data-driven" pour une grande classe de séries chronologiques, qui comprend les processus ARMA ou AR(∞), ainsi que GARCH ou ARCH(∞), APARCH et bien d'autres. Nous abordons la question difficile de la conception de critères adaptatifs jouissant de la propriété de forte consistance. Lorsque les observations proviennent de l'un des modèles susmentionnés, les nouveaux critères retrouvent le vrai modèle presque sûrement parmi une famille finie de modèles candidats, ceci quand la taille des données croit. Les critères proposés sont basés sur la minimisation d'un contraste pénalisé similaire au critère de Hannan et Quinn et impliquent un terme de pénalité connu pour la plupart des modéles de séries chronologiques classiques et pour les modèles plus complexes, ce terme est inconnu mais peut être calibré uniquemement par les données.

Mots-clés. Séries chronologiques, sélection de modèles, consistance, data-driven, critère HQ.

Abstract. This paper aims to study data driven model selection criteria for a large class of time series, which includes ARMA or AR(∞) processes, as well as GARCH or ARCH(∞), APARCH and many others processes. We tackled the challenging issue of designing adaptive criteria which enjoys the strong consistency property. When the observations are generated from one of the aforementioned models, the new criteria, select the true model almost surely asymptotically. The proposed criteria are based on the minimization of a penalized contrast akin to the Hannan and Quinn's criterion and then involved a term which is known for most classical time series models and for more complex models, this term can be data driven calibrated.

Keywords. Time series, Model selection, consistency, data driven, HQ criterion.

1 INTRODUCTION

A common solution in model selection is to choose the model, minimizing a penalized based criterion which is the sum of two terms: the first one is the empirical risk (least squares, likelihood) that measures the goodness of fit and the second one is an increasing function of the complexity which aims to penalize large models and control the bias. Therefore a challenging task when designing a penalized criterion is the specification of the penalty term. Considering leading model selection criteria (BIC, AIC, Cp, HQ to name a few), one can see that the penalty term is a product of the model dimension with a sequence which is specific to the criteria. Indeed, a criterion is designed according to

the goal one would like to achieve. The classical properties for model selection criteria include *consistency*, *efficiency* (oracle inequality, asymptotic optimality), *adaptative in the minimax sense*.

In this paper, we focus on consistency property which aims at identifying the data generating process with high probability or almost surely. Hence, it requires the assumption whereby there exists a true model in the set of competitive models and the goal is to select this with probability approaches one as the sample size tends to infinity.

Compare to HQ penalty, the BIC penalty does not have the slowest rate of increase and then it can very often choose very simple models possible wrongs for small samples Hannan and Quinn (1979). Moreover, the HQ criterion has been derived for linear time series: AR models in Hannan and Quinn (1979), ARMA models in Hannan (1979) and Hannan and Deistler (2012). Is the HQ penalty still strongly consistent for heteroscedastic nonlinear models such as GARCH, APARCH or ARMA-GARCH? And what about a general class including linear and non linear models as well?

This is the issue we want to address in this paper for a general class of times series called affine causal.
The main contribution of this paper is the generalization of the HQ criterion to affine causal class: we provide a minimal multiplicative penalty term c_{min} so that all penalties of the form $2\,c\,\log\log n\,D_m$ with $c \geq c_{min}$ ensure the strong consistency property for affine causal models (D_m denotes the size of the model m). The minimal constant is known for classical models (ARMA, GARCH or APARCH type) but for the most complex ones, it can be data-driven calibrated using slope heuristic (see Birgé and Massart (2007a), Arlot and Massart (2009)). However, the theoretical validity of this heuristic, especially for time series, is an open research topic.

2 Affine Causal Class

Class $\mathcal{AC}(M, f)$: A process $X = (X_t)_{t\in\mathbb{Z}}$ belongs to $\mathcal{AC}(M, f)$ if it satisfies:

$$X_t = M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t + f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \text{ for any } t \in \mathbb{Z}. \tag{2.1}$$

where $(\xi_t)_{t\in T}$ is a sequence of zero-mean independent identically distributed random vectors (i.i.d.r.v) satisfying $\mathbb{E}(|\xi_0|^r) < \infty$ with $r \geq 1$ and $M,\ f : \mathbb{R}^\infty \to \mathbb{R}$ are two measurable functions. The function f can be considered as the conditional mean of the process while M^2 can be seen as the conditional variance which will allow to take into account a possible heteroscedasticity.

For instance,

- if $M\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sigma$ and $f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sum_{i=1}^{\infty}\phi_i X_{t-i}$, then $(X_t)_{t\in\mathbb{Z}}$ is an AR(∞) process;
- if $M\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sqrt{a_0 + a_1X_{t-1}^2 + \cdots + a_pX_{t-p}^2}$ and $f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = 0$, then $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process.

Note that, numerous classical time series models such as ARMA(p, q), GARCH(p, q), ARMA(p, q)-GARCH(p, q) or APARCH(δ, p, q) processes belongs to $\mathcal{AC}(M, f)$.

The study of this type of process more often requires the classical regularity conditions on the functions M and f, which are not restrictive at all and remain valid in various time serie models. Let us recall these conditions for $\Psi_\theta = f_\theta$ or M_θ and Θ a compact set.

Hypothesis A(Ψ_θ, Θ): *Assume that* $\|\Psi_\theta(0)\|_\Theta < \infty$ *and there exists a sequence of non-negative real numbers* $\big(\alpha_k(\Psi_\theta, \Theta)\big)_{k\geq 1}$ *such that* $\sum_{k=1}^{\infty} \alpha_k(\Psi_\theta, \Theta) < \infty$ *satisfying:*

$$\|\Psi_\theta(x) - \Psi_\theta(y)\|_\Theta \leq \sum_{k=1}^{\infty} \alpha_k(\Psi_\theta, \Theta)|x_k - y_k| \text{ for all } x, y \in \mathbb{R}^\infty.$$

In addition, if the noise ξ_0 admits r-order moments (for $r \geq 1$), let us define:

$$\Theta(r) = \Big\{\theta \in \mathbb{R}^d,\ A(f_\theta, \{\theta\}) \text{ and } A(M_\theta, \{\theta\}) \text{ hold with } \sum_{k=1}^{\infty} \alpha_k(f_\theta, \{\theta\}) + \|\xi_0\|_r \sum_{k=1}^{\infty} \alpha_k(M_\theta, \{\theta\}) < 1\Big\}. \quad (2.2)$$

Under this assumption, Doukhan and Wintenberger (2008) showed that there exists a stationary and ergodic solution to (2.1) with r-order moment for any $\theta \in \Theta(r)$. Moreover, Bardet and Wintenberger (2008) studied the consistency and the asymptotic normality of the QMLE of θ^* for $\mathcal{AC}(M_{\theta^*}, f_{\theta^*})$.

3 Model Selection Procedure

Let assume $(X_1, \ldots, X_n)$ be a trajectory of a stationary affine causal process $m^* := \mathcal{AC}(M_{\theta^*}, f_{\theta^*})$, where θ^* is unknown. The goal of the consistency property is to come up with this true model given a set of candidate model $\mathcal{M}$ such that $m^* \in \mathcal{M}$.

A D_m-dimensionnal model $m \in \mathcal{M}$ can be viewed as a set of causal functions (M_θ, f_θ) with $\theta \in \Theta(m) \subset \mathbb{R}^{D_m}$. $\Theta(m)$ is the parameter set of the model m.
The consistency property will be studied using quasi likelihood estimation since assumption on the distribution of the noise is not required.
The Gaussian quasi log-likelihood is derived from the conditional (with respect to the filtration $\sigma\big(X_t, t \leq 0\big)$ log-likelihood of $(X_1, \ldots, X_n)$ when (ξ_t) is supposed to be a Gaussian standard white noise. From (2.1), one deduce that the log density of X_t given $\sigma\big(X_i, i < t\big)$ is

$$-\frac{1}{2}\Big[\frac{(X_t - f^t_{\theta^*})^2}{H^t_{\theta^*}} + \log(H^t_{\theta^*})\Big].$$

Therefore the conditional log density of $(X_1, \ldots, X_n)$ given $\sigma\big(X_t, t \leq 0\big)$ is

$$-\frac{1}{2}\sum_{t=1}^{n}\Big[\frac{(X_t - f^t_{\theta^*})^2}{H^t_{\theta^*}} + \log(H^t_{\theta^*})\Big].$$

From now on, we drop the Gaussian assumption of the noise. The conditional log-density inspires to define for all $\theta \in \Theta$

$$L_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} q_t(\theta)\ ,\ \text{with } q_t(\theta) := \frac{(X_t - f^t_\theta)^2}{H^t_\theta} + \log(H^t_\theta) \quad (3.1)$$

where $f_\theta^t := f_\theta(X_{t-1}, X_{t-2}, \cdots)$, $M_\theta^t := M_\theta(X_{t-1}, X_{t-2}, \cdots)$ and $H_\theta^t = \left(M_\theta^t\right)^2$. The quasi likelihood function L_n is not computable since it depends on the past $(X_{-j})_{j\in\mathbb{N}}$ that is unknown. However, the sequence $(L_n(.))_n$ enjoys very nice asymptotic properties such as the Uniform Law of Large Numbers (see Bardet and Wintenberger (2008)).
Therefore, we consider an observable approximation of L_n denoted $\widehat{L}_n$ and define as follows

$$\widehat{L}_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} \widehat{q}_t(\theta) \ , \text{ with } \widehat{q}_t(\theta) := \frac{(X_t - \widehat{f}_\theta^t)^2}{\widehat{H}_\theta^t} + \log(\widehat{H}_\theta^t) \tag{3.2}$$

where $\widehat{f}_\theta^t := f_\theta(X_{t-1}, X_{t-2}, \ldots, X_1, 0, \ldots, 0)$, $\widehat{M}_\theta^t := M_\theta(X_{t-1}, X_{t-2}, \ldots, X_1, 0, \ldots, 0)$ and $\widehat{H}_\theta^t = \left(\widehat{M}_\theta^t\right)^2$.

Let $\mathcal{M}$ a finite family of candidate models containing the true one m^*. According to Proposition 1 in Bardet and al. (2020), all these models can be included into a big one with parameter space Θ. For each specific model $m \in \mathcal{M}$, we define the Gaussian QMLE $\widehat{\theta}(m)$ as

$$\widehat{\theta}(m) = \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\, \widehat{L}_n(\theta). \tag{3.3}$$

To select the true model $m^* \in \mathcal{M}$, we consider a penalized contrast $\widehat{C}$

$$\widehat{C}(m) = -2\,\widehat{L}_n\big(\widehat{\theta}(m)\big) + \kappa_n(m) \quad \text{for all } m \in \mathcal{M},$$

ensuring a trade-off between -2 times the maximized quasi log-likelihood, which decreases with the size of the model, and a penalty increasing with the size of the model. Therefore, the selection procedure consists to choose as an estimator of m^*, the model which minimizes $\widehat{C}$ over the family $\mathcal{M}$, that is:

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\, \widehat{C}(m). \tag{3.4}$$

There exist several possible choices of κ_n including

- $\kappa_n(m) = 2\,c\,D_m \log\log n$ with $c > 1$, we retrieve the HQ criterion (see Hannan and Quinn (1979));
- $\kappa_n = D_m \log n$, $\widehat{C}$ yields to BIC criterion (Schwarz (1978)) ;
- $\kappa_n = 2\,D_m$, $\widehat{C}$ is the AIC criterion (Akaike (1973)).

4 Assumptions and Consistency Result

Some mild conditions will be required to prove the consistency of the considered model selection criteria.

Assumption A1: *For all* $\theta, \theta' \in \Theta_m$, $(f_\theta^0 = f_{\theta'}^0)$ *and* $(M_\theta^0 = M_{\theta'}^0) \implies \theta = \theta'$.

Assumption A2: *One of the families* $(\partial f_\theta^t/\partial\theta^{(i)})_{1\le i\le D_{m^*}}$ *or* $(\partial H_\theta^t/\partial\theta^{(i)})_{1\le i\le D_{m^*}}$ *is a.e. linearly independent.*

Assumption A3: $\exists \underline{h} > 0$ *such that* $\inf_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ *for all* $x \in \mathbb{R}^\infty$.

Assumption A4: $\mathbb{E}[\xi_0^8] < \infty$.

We assume a suitable relation between the Fisher Information matrix $G(\theta_m^*)$ and the limiting Hessian matrix of the log-likelihood $F(\theta_m^*)$ defined as follows

$$\big(F(\theta_m^*)\big)_{i,j} = \mathbb{E}\Big[\frac{\partial^2 q_0(\theta_m^*)}{\partial\theta_i\partial\theta_j}\Big] \quad \text{and} \quad (G(\theta_m^*))_{i,j} = \mathbb{E}\Big[\frac{\partial q_0(\theta_m^*)}{\partial\theta_i}\frac{\partial q_0(\theta_m^*)}{\partial\theta_j}\Big],$$

with $\theta_m^* := (\theta^*, 0, \ldots, 0)^\top \in \Theta(m)$.

Assumption A5: There exist absolutes constants α_1 and α_2 such that for any $m \in \mathcal{M}$ verifying $m^* \subset m$,

$$\mathbf{1}_m^\top \Sigma_{\theta_m^*} \mathbf{1}_m = \alpha_1\, D_m^1 + \alpha_2\, D_m^2 \tag{4.1}$$

where D_m^1 and D_m^2 are two integers such that $D_m^1 + D_m^2 = D_m$, $\quad \mathbf{1}_m := (1, 1, \ldots, 1)^\top \in \mathbb{R}^{D_m}$, $\Sigma_{\theta_m^*} := G(\theta_m^*)^{1/2} F(\theta_m^*)^{-1} G(\theta_m^*)^{1/2}$.
For most classical affine causal models, **A5** is verified (see Proposition 2). However, for more complex models such as ARMA-GARCH with $\mu_4 \neq 3$, $\Sigma_{\theta_m^*}$ is hard to handle.

In this framework (see Bardet and Wintenberger (2009), Bardet and al. (2020)), it is classical to not take into account long memory process, that is to assume that the Lispchitz coefficients satisfy the following conditions

Assumption A6 $\qquad \alpha_j(f_\theta)+\alpha_j(M_\theta)+\alpha_j(\partial_\theta f_\theta)+\alpha_j(\partial_\theta M_\theta) = O(j^{-\gamma}) \quad$ with $\quad \gamma > 3/2.$

The following Proposition suggests the existence of a term that will be the keystone of this work.

Proposition 1. *Let* m^* *any affine causal model. For any model* m *with* $\theta_m^* \in \widetilde{\overset{\circ}{\Theta}(m)}$, *and under* ***A1-A6****, there exist* α_1, α_2, D_m^1, D_m^2 *such that*

$$\limsup_{n\to\infty} \frac{L_n\big(\widehat{\theta}(m)\big) - L_n(\theta_m^*)}{2\,\log\log n} = \frac{1}{4}\big(\alpha_1\, D_m^1 + \alpha_2\, D_m^2\big) \quad a.s. \tag{4.2}$$

For every $m \in \mathcal{M}$, let us denote by $c_{min}(m)$ the following term that will be used several times

$$c_{min}(m) := \frac{1}{4}\big(\alpha_1\, D_m^1 + \alpha_2\, D_m^2\big) \tag{4.3}$$

Now we state a result which provides the values of both α_1 and α_2 for most classical affine causal models.

Proposition 2. *Under the assumptions and notation of Proposition 1, we have*

- *If* $\mu_4 = \mathbb{E}[\xi_0^4] = 3$ *(for instance for Gaussian noise), then* $\alpha_1 = 2$, $\alpha_2 = 2$ *and* $c_{min}(m) = \frac{1}{2}\, D_m$*;*
- *If the parameter* θ *identifying an affine causal model* $X_t = M_\theta^t\, \xi_t + f_\theta^t$ *can be decomposed as* $\theta = (\theta_1, \theta_2)'$ *with* $f_\theta^t = \widetilde{f}_{\theta_1}^t$ *and* $M_\theta^t = \widetilde{M}_{\theta_2}^t$, *then* $\alpha_1 = 2$, $\alpha_2 = \mu_4 - 1$ *and*

$$c_{min}(m) = \frac{1}{2}\, D_m^1 + \frac{\mu_4 - 1}{4}\, D_m^2$$

The second configuration in Proposition 2 includes classical time series

- GARCH(p,q), APARCH(δ,p,q) type models and related ones, $c_{min}(m)=\frac{\mu_4-1}{4}D_m$;
- ARMA(p,q) models, $c_{min}(m)=\frac{D_m}{2}$ if the variance of the noise is known and $c_{min}(m)=\frac{D_m-1}{2}+\frac{\mu_4-1}{4}$ otherwise.

We can now state the main result of this paper.

Theorem 4.1. *Let $(X_1,\dots,X_n)$ be an observed trajectory of an affine causal process X belonging to $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$ where θ^* is an unknown vector belonging to $\Theta(r)\subset\mathbb{R}^{D_{m^*}}$. Let also $\mathcal{M}$ be a finite family of candidate models such that $m^*\in\mathcal{M}$. If assumptions **A1-A6** hold, there exist α_1, α_2, and a minimal constant $c_{min}:=\max\left(\frac{\alpha_1}{4},\frac{\alpha_2}{4}\right)$ such that*

for any $\kappa_n(m)=2\,c\,D_m\,\log\log n$ with

$$c\geq c_{min} \tag{4.4}$$

it holds for the selected model $\widehat{m}$ according to (3.4)

$$\widehat{m}\xrightarrow[n\to\infty]{a.s.}m^*. \tag{4.5}$$

Remark 1.
1. *For classical configurations as seen in Proposition 2, this result gives a generalization of Hannan and Quinn criterion.*
2. *For more complex models, the values of α_1 and α_2 are unknowns (at least until a better relationship between matrix $F(\theta_m^*)$ and $G(\theta_m^*)$ is found) and so c_{min} is also unknown. In these cases, we propose to use adaptive methods such as slope heuristic algorithm or dimension jump Arlot and Massart (2009) to calibrate c_{min}.*

Bibliography

Arlot, S., and Massart, P. Data-driven calibration of penalties for least-squares regression. Journal of Machine learning research 10 (2009), 245–279.

Akaike, H. Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd international symposium on information, Akademiai Kiado, Budapest (1973).

Bardet, J.-M., and Wintenberger, O. Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. The Annals of Statistics 37, 5B (2009), 2730–2759.

Bardet, J.-M., Kamila, K., and Kengne, W. Consistent model selection criteria and goodness-of-fit test for common time series models. Electronic Journal of Statistics 14, 1 (2020), 2009–2052.

Doukhan, P., and Wintenberger, O. Weakly dependent chains with infinite memory. Stochastic Processes and their Applications 118, 11 (2008), 1997–2013.

Hannan, E. J., and Quinn, B. G. The determination of the order of an autoregression. Journal of the Royal Statistical Society: Series B (Methodological) 41, 2 (1979), 190–195

Hannan, E. The estimation of the order of an arma process. The Annals of Statistics 8, 5 (1980), 1071–1081.

Hannan, E. J., and Deistler, M. The statistical theory of linear systems. SIAM, 2012.

Schwarz, G. Estimating the dimension of a model. The Annals of Statistics 6 (1978), 461–464.

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. Probability theory and related fields, 138(1-2):33–73, 2007a.

Analyse de sensibilité globale de modèles stochastiques à sorties fonctionnelles

Henri Mermoz KOUYE [1] & Gildas MAZO[2] & Clémentine PRIEUR[3] & Elisabeta VERGU[4]

[1] *MaIAGE, INRAE, Université Paris Saclay, 78350 Jouy-en-Josas*
E-mail : henri-mermoz.kouye@inrae.fr
[2] *MaIAGE, INRAE, Université Paris Saclay, 78350 Jouy-en-Josas*
E-mail : gildas.mazo@inrae.fr
[3] *Laboratoire Jean Kuntzmann, Université Grenoble Alpes, 38041 Grenoble cedex 9*
E-mail : clementine.prieur@univ-grenoble-alpes.fr
[4] *MaIAGE, INRAE, Université Paris Saclay, 78350 Jouy-en-Josas*
E-mail : elisabeta.vergu@inrae.fr

Résumé. Les modèles stochastiques sont de plus en plus utilisés dans divers domaines (épidémiologie, biologie, physique, etc.) et peuvent être représentés par des processus stochastiques. L'utilisation de ce type de processus pour décrire un phénomène dynamique permet d'intégrer à la description un aléa intrinsèque. La notion d'analyse de sensibilité globale (GSA) pour des modèles stochastiques est délicate, en particulier lorsque l'aléa intrinsèque est considéré comme un bruit sur des quantités d'intérêts spécifiques (QoI). L'objectif de notre travail est de proposer une stratégie générique d'analyse de sensibilité (AS) sur une classe de modèles stochastiques utilisés en épidémiologie pour décrire des dynamiques épidémiques. Notre stratégie vise à séparer les deux sources de variabilité, à savoir les paramètres incertains et l'aléa intrinsèque en écrivant la quantité QoI sous la forme d'une fonction déterministe des paramètres incertains X et de l'aléa intrinsèque Z codé par un vecteur aléatoire, avec X et Z indépendants. Pour cela, deux approches sont introduites : la construction de Sellke (1983) et la représentation de Kurtz (1982,1986). Ces deux approches permettent non seulement d'estimer les indices de sensibilité des paramètres d'entrée mais aussi de mesurer l'influence de l'aléa intrinsèque sur la variabilité globale de la QoI étudiée. De plus, la construction de Sellke peut être utilisée pour étendre la méthodologie au cadre non markovien. Partant de là, différents indices de sensibilité seront ensuite estimés. Ces approches seront appliquées à un modèle simplifié de propagation de SARS-CoV-2 (Knock, 2021) pour identifier les paramètres importants dans la dynamique de l'épidémie.

Mots-clés. Chaînes de Markov à temps continu, processus de comptage, modèles épidémiologiques, analyse de sensibilité globale, décompositions basées sur les noyaux.

Abstract. Stochastic models are increasingly used in various fields (epidemiology, biology, physics etc) to describe different phenomena. Indeed, many phenomena can be described by stochastic processes, hence including an intrinsic randomness. Performing

global sensitivity analysis (GSA) for such models is challenging, in particular if the intrinsic randomness of the system is considered as a noise on specific quantities of interests (QoIs). The objective of our work is to propose a strategy to perform GSA on a generic class of stochastic models used in epidemiology to describe epidemic dynamics. Our strategy aims at separating the two sources of variability, namely parameters uncertainty and intrinsic randomness by putting the model functional QoI as a deterministic function of uncertain parameters X and of the intrinsic randomness coded in a random vector Z such as X and Z are independent. For this purpose, two approaches are introduced : Sellke construction (1983) and Kurtz representation (1982,1986). These two approaches allow not only to estimate sensitivity indices of the input parameters but also to measure the influence of the intrinsic randomness on the global variability of the QoI under study. Furthermore, Sellke construction can be used to extend our approach to the non-markovian framework. Then, we put ourselves in various frameworks of sensitivity analysis and implement different sensitivity measures. These approaches will be applied to a simplified model of SARS-CoV-2 diffusion (Knock, 2021) in order to identify key parameters of the epidemic dynamics.

Keywords. continuous time Markov chains, counting processes, epidemic models, globale sensitivity analysis, kernel-based decompositions.

Introduction

La modélisation mathématique est de plus en plus utilisée dans des domaines variés pour décrire des phénomènes divers aussi bien à petite qu'à grande échelle. Une partie importante de ces modèles est basée sur des processus aléatoires (processus de sauts, processus dérivant des équations différentielles stochastiques ...). Ainsi, non seulement les sorties de ces modèles ou les quantités d'intérêt (QoI) peuvent être des fonctions, des trajectoires au cours du temps, mais aussi elles sont intrinsèquement aléatoires : deux appels du modèle évalué à une même entrée peuvent donner lieu à des résultats différents. Selon le phénomène étudié, ces modèles peuvent être complexes (non-linéaires, espace d'états de dimension élevée, etc.) et dépendre d'un nombre important de paramètres d'entrée (encore désignés sous la terminologie entrées du modèle) potentiellement mal connus. Dans ce contexte, l'analyse de sensibilité (AS) est un outil important à la fois pour mieux appréhender un phénomène et pour éventuellement réduire la dimension de l'espace des paramètres.

Etat de l'art

Plusieurs travaux ont été menés pour mettre au point des méthodes d'AS pour des modèles stochastiques. D'une part Hart et al (2017) et d'autre part Mazo (2017) ont

proposé un cadre d'AS des modèles à sortie scalaire en définissant des indices de sensibilité inspirés des indices de Sobol' (1990) et en proposant des procédures d'estimation efficace. De plus, une approche d'analyse de sensibilité globale a été introduite par Etoré et al (2020) pour les modèles stochastiques décrits par des équations différentielles stochastiques. En partant des modèles issus de la physico-chimie, Navarro et al (2016) ont utilisé la représentation de Kurtz (1982,1986) pour obtenir une décomposition de Sobol-Hoeffding de la variance pour des modèles à sortie scalaire basés sur une classe de chaînes de Markov. D'autres approches reposent sur la méta-modélisation qui consiste à émuler le comportement du modèle original avec un modèle moins complexe et pour lequel l'analyse de sensibilité est plus simple à mettre en oeuvre (Zhu, 2021). Toutes ces approches et les différents cadres utilisés révèlent toute la difficulté de maîtriser et gérer l'aléa intrinsèque dans l'AS de ces modèles. En pratique, cette difficulté se traduit par un coût élevé en temps de calcul ou en précision des estimations.

Résultats

Nos études portent sur l'analyse de sensibilité globale des modèles stochastiques basés sur des processus de sauts, en l'occurrence des modèles compartimentaux fréquemment utilisés en épidémiologie pour représenter des dynamiques. Notre approche pour ce type de modèles est de séparer les deux sources de variabilité : l'incertitude sur les paramètres inconnus et l'aléa intrinsèque. Nous montrons en effet que sous certaines hypothèses les modèles compartimentaux considérés s'écrivent sous la forme :

$$Y = f(X, Z), \tag{1}$$

avec f une fonction déterministe, X le vecteur aléatoire des paramètres incertains, Z un vecteur aléatoire de loi connue codant l'aléa intrinsèque et où X et Z sont indépendants. La forme (1) présente principalement deux avantages. En premier lieu, elle permet de se placer dans un cadre d'AS très bien étudié qui est celui des modèles déterministes, car en échantillonnant aussi l'aléa intrinsèque, les réponses du modèle sont fixes dès qu'on fige les valeurs des paramètres et de l'aléa. En second lieu, la connaissance de la distribution de l'aléa intrinsèque permet de quantifier sa contribution à la variance globale. Pour obtenir la représentation (1), nous utilisons deux approches : la représentation de Kurtz (dans le cas de modèles markoviens) et la construction de Sellke (dans les cas de modèles markoviens et non markoviens). La représentation de Kurtz sert à écrire des chaînes de Markov à temps continu en fonction de processus de Poisson d'intensité 1. Alors, il devient possible d'écrire toutes les sorties (éventuellement fonctionnelles) comme fonction des paramètres et d'un nombre fini de processus de Poisson. Le vecteur des processus de Poisson constitue l'aléa intrinsèque Z. Quant à la construction de Sellke, elle consiste à redéfinir les mécanismes de sauts du modèle de façon à ce qu'ils dépendent de l'évolution d'une ou plusieurs fonctions cumulatives croissantes. Les sorties s'obtiennent ainsi comme des fonctions des paramètres et d'un vecteur aléatoire de longueur finie constitué de variables exponentielles et uniformes ou multinomiales.

Nous employons ces deux techniques pour réaliser l'analyse de sensibilité globale d'une version simplifiée d'un modèle de propagation de l'épidémie de SARS-CoV-2 proposé dans Knock et al (2021). Le modèle étudié est un modèle à sept compartiments correspondant chacun à un état. Un individu peut être susceptible (S), exposé (E), infectieux asymptomatique (IA), infectieux symptomatique (IS), hospitalisé (H), immunisé (R) ou décédé (D). Le modèle est basé sur des processus de saut dont les transitions d'états, schématisées dans la figure (fig. 1) dépendent de huit paramètres inconnus : le taux de transmission β (caractérisant la transition $S \to E$), la durée moyenne d'incubation $1/\delta$ (qui caractérise les transitions $E \to IA$ et $E \to IS$), la durée moyenne d'infection des asymptomatiques $1/\mu_A$ (transition $IA \to R$), la durée moyenne d'infection des symptomatiques $1/\mu_S$ (transitions $IS \to R$, $IS \to H$ et $IS \to D$), la durée moyenne d'hospitalisation $1/\gamma_H$ (transitions $H \to R$ et $H \to D$), la probabilité p_S pour un individu exposé de présenter des symptômes, la probabilité $p = (p_1, p_2, p_3)$ pour un individu symptomatique de guérir, d'être hospitalisé ou de décéder, la probabilité p_H pour un individu hospitalisé de guérir. Nous nous intéressons à l'influence de ces paramètres sur la dynamique du nombre d'individus infectieux asymptomatiques, symptomatiques, hospitalisés et décédés (voir fig. 2 pour un exemple de ces dynamiques). Dans ce travail, nous proposons d'estimer pour ces quantités d'intérêts fonctionnelles dépendant des paramètres incertains et de l'aléa intrinsèque, des indices de Sobol' dynamiques, agrégés (Lamboni, 2011) et des indices de type HSIC (Da Veiga, 2021) en partant de la représentation (1).

Figures

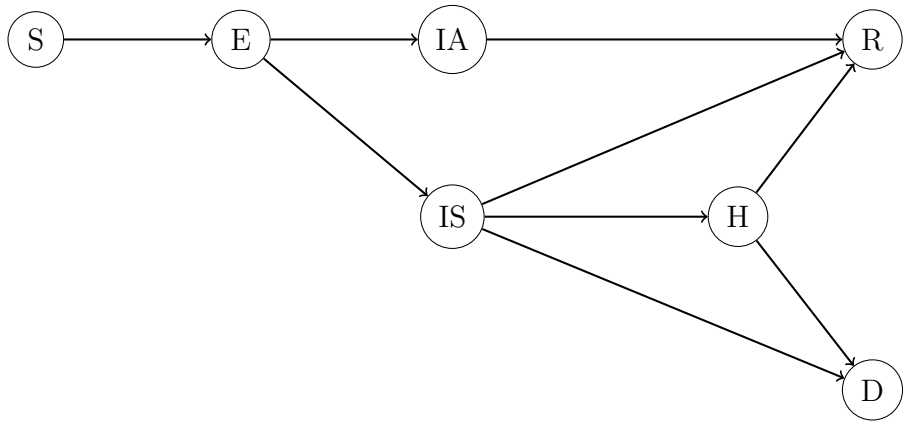


FIGURE 1 – Modèle compartimental de propagation du SARS-CoV-2 compartant 7 états (S,E,IA,IS,H,R,D) et 8 paramètres ($\beta, \delta, \mu_A, \mu_S, \gamma_H, p_S, p, p_H$).

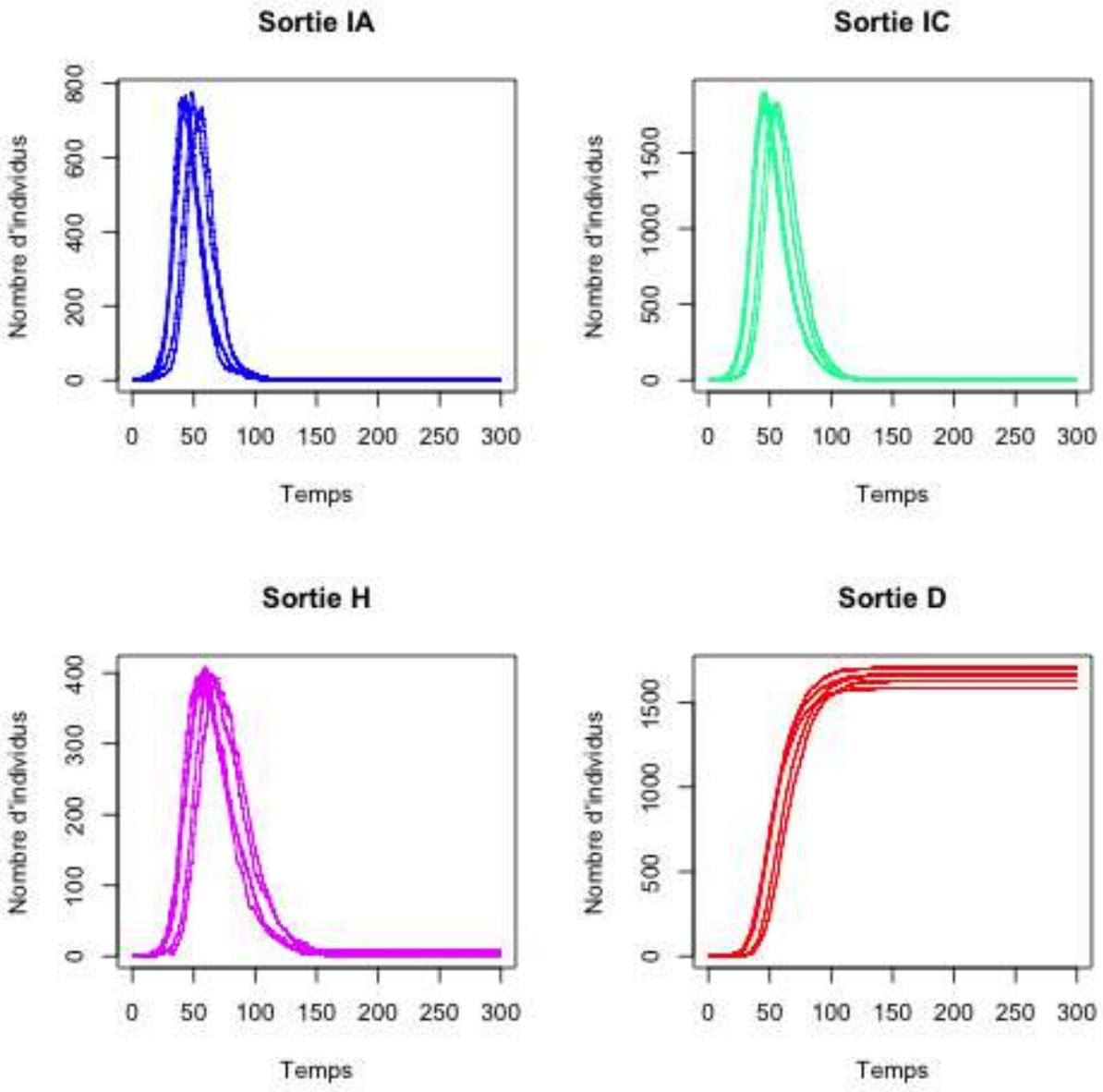


FIGURE 2 – Exemples de dynamiques d'évolution des nombres d'individus asymptomatiques (IA), symptomatiques (IC), hospitalisés (H) et décédés (D). Simulations par la construction de Sellke de 5 trajectoires pour des paramètres d'entrés $\beta = 1$, $1/\delta = 9$, $1/\mu_A = 5, 1/\mu_C = 10, 1/\gamma_H = 15, p_S = 60\%, p = (60\%, 20\%, 20\%), p_H = 60\%$ avec un nombre initial de susceptibles égal à 10 000 et un nombre initial d'individus exposés égal à 5.

Bibliographie

Hart, J. L. and Alexanderian, A. and Gremaud, P. A.(2017). Efficient Computation of Sobol' Indices for Stochastic Models, *SIAM Journal on Scientific Computing*,39,pp. A1514-A1530.

I.M. Sobol'.(1990), Sensitivity estimates for nonlinear mathematical models,*Matematicheskoe Modelirovanie 2*,pp. 112–118.

Mazo, G. (2021). Global sensitivity indices, estimators and tradeoff between explorations and repetitions for some stochastic models.,*working paper or preprint hal-02113448*.

Étoré, P. and Prieur, C. and Pham, D. K. and Li, L.(2020). Global Sensitivity Analysis for Models Described by Stochastic Differential Equations,*Methodology and Computing in Applied Probability*,pp.803-831.

Kurtz, T. G.(1982).Representation and approximation of counting processes, Advances in Filtering and Optimal Stochastic Control,*Springer Berlin Heidelberg*,pp.177-191.

Navarro Jimenez,M. and Le Maître,O. P. and Knio,O. M. (2016). Global sensitivity analysis in stochastic simulators of uncertain reaction networks,*The Journal of Chemical Physics*, pp. 145(24) :244106.

Ethier, Stewart N. and Kurtz, Thomas G.(1986). Markov processes : characterization and convergence,*Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics*.

Sellke, Thomas. (1983). On the asymptotic distribution of the size of a stochastic epidemic, *Journal of Applied Probability*, 20,pp.390–394.

Knock, Edward S. and al (2021). The 2020 SARS-CoV-2 epidemic in England : key epidemiological drivers and impact of interventions,*medRxiv*, doi = 10.1101/2021.01.11.21249564.

Matieyendou, L. and Hervé, M. and David, M.(2011).Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models,*Reliability Engineering and System Safety*,96,pp.450 - 459.

Da Veiga, Sébastien.(2021). Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis,*working paper or preprint hal-03108628*.

X. Zhu and B. Sudret.(2021). Global sensitivity analysis for stochastic simulators based on generalized lambda surrogate models,*arXiv*, https ://arxiv.org/abs/2005.01309.

Efficient Bayesian data assimilation via inverse regression

Benoit KUGLER [1] & Florence FORBES [1] & Sylvain DOUTE [2] & Michel GAY [3]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France (firstname.lastname@inria.fr)*
[2] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France (sylvain.doute@univ-grenoble-alpes.fr)*
[3] *Grenoble Image sPeech Signal and Automatics Lab, Grenoble 38000 Grenoble, France (michel.gay@gipsa-lab.grenoble-inp.fr)*

Résumé. On propose une approche bayésienne pour résoudre un problème d'assimilation de données. Dans un premier temps, le modèle direct est approché par un modèle paramétrique inversible. Dans un deuxième temps, l'information a-priori est intégrée. Cette division en deux étapes permet de traiter efficacement un nombre important d'inversions. La méthode est illustrée sur une étude du manteau neigeux, utilisant un modèle de rétro-diffusion électro-magnétique.

Mots-clés. Assimilation de données, problème inverse, régression, apprentissage statistique

Abstract.
We propose a Bayesian approach to data assimilation problems, involving two steps. We first approximate the forward physical model with a parametric invertible model, and we then use its properties to leverage the availaibility of a priori information. This approach is particularly suitable when a large number of inversions has to be performed. We illustrate the proposed methodology on a multilayer snowpack model.

Keywords. Data assimilation, inverse problem, regression, statistical learning

1 Introduction

A data assimilation task aims at retrieving unknown parameters, denoted by $\mathbf{x}$, from observations $\mathbf{y}$ and an initial guess on the parameters $\mathbf{x}_0$. The observations and the parameters are linked by a forward model, denoted by F. This problem is similar to inverse problems but differs in the sense that the number of observations $\mathbf{y}$ is much smaller than the number of parameters so that the observations $\mathbf{y}$ alone are not enough to predict the parameters. It is then crucial to make full use of an initial guess of the parameters $\mathbf{x}_0$ to avoid an ill-posed problem. One way to formalize this problem is to

adopt a Bayesian formulation. Our problem is modeled considering two random variables $\mathbf{X} \in \mathbb{R}^L$ and $\mathbf{Y} \in \mathbb{R}^D$, linked by the relation

$$Y = F(X) + \varepsilon \tag{1}$$

where ε is a centered Gaussian noise, with variance $\mathbf{\Sigma}$, accounting for the measurement and model uncertainties. We then account for an initial guess $\mathbf{x}_0$ with a prior density on $\mathbf{X}$, for example the product of a Gaussian distribution with mean $\mathbf{x}_0$ and a variance $\mathbf{\Gamma}_0$ and of a uniform distribution on the parameters range, denoted by $\mathcal{U}_\mathcal{P}$. According to Bayes' rule, the posterior distribution has then the following form:

$$p^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \propto \mathcal{U}_\mathcal{P}(\mathbf{x})\mathcal{N}\left(\mathbf{x}; \mathbf{x}_0, \mathbf{\Gamma}_0\right)\mathcal{N}\left(F(\mathbf{x}); \mathbf{y}, \mathbf{\Sigma}\right) \ .$$

The choice of $\mathbf{\Gamma}_0$ and $\mathbf{\Sigma}$ is crucial. Taking $\mathbf{\Sigma} = 0$ boils down to solve the inverse problem alone, without taking into account prior information. In contrast, taking $\mathbf{\Gamma}_0 = 0$ just yields a dirac centered at $\mathbf{x}_0$, without exploiting the measurements. When looking at the maximum a posteriori (MAP) solution, it comes

$$\hat{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x}\in P} ||\mathbf{x} - \mathbf{x}_0||_{\mathbf{\Gamma}_0} + ||\mathbf{y} - F(\mathbf{x})||_{\mathbf{\Sigma}} \ ,$$

where $||\cdot||_{\mathbf{\Sigma}}$ denotes the Mahalanobis distance. This is a well known consequence of assuming Gaussian distributions for the forward and prior models. In this work, we propose to go beyond the Gaussian assumption by learning the underlying relation between $\mathbf{X}$ and $\mathbf{Y}$, using a regression approach. We first introduce the statistical model and show how it can be used in an assimilation problem. We then illustrate the method on a realistic example in remote sensing.

2 Efficient assimilation via regression

We propose to use a two-steps approach: first, we consider the problem without prior information, and we learn the underlying relation between $\mathbf{X}$ and $\mathbf{Y}$, using the so-called Gaussian Locally-Linear Mapping model (GLLiM) ([Deleforge et al., 2015]). Then, we adapt the model to take into account prior information.

2.1 Learning an invertible approximation of the forward model

In this first step, the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is approximated by a Gaussian Locally-Linear Mapping model (GLLiM) which builds upon Gaussian mixture models to capture non linear relationships ([Deleforge et al., 2015]). A latent variable $Z \in \{1, \ldots, K\}$ is introduced to model $\mathbf{Y}$ as piece-wise affine transformation of $\mathbf{X}$:

$$\mathbf{Y} \ = \ \sum_{k=1}^{K} \mathbb{I}_{\{Z=k\}}(\boldsymbol{A}_k\mathbf{X} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k) \tag{2}$$

where $\mathbb{I}$ is the indicator function, $\boldsymbol{A}_k$ a $D \times L$ matrix and $\boldsymbol{b}_k$ a vector of $\mathbb{R}^D$ that define an affine transformation. Variable $\boldsymbol{\epsilon}_k$ corresponds to an error term which is assumed to be zero-mean and not correlated with $\mathbf{X}$ capturing both the observation noise and the reconstruction error due to the affine approximation.

In order to keep the posterior tractable, we assume that $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$ and $\mathbf{X}$ is a mixture of K Gaussians : $p(\mathbf{x}|Z = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ and $p(Z = k) = \pi_k$. The GLLiM model is thus characterized by the parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1:K}$

This model can be learned from a training set using an EM algorithm. More specifically, the training set $(\mathbf{x}_n, \mathbf{y}_n)_{n=1..N}$ is simulated such that $\mathbf{x}_n$ are realizations of the prior $\mathcal{U}_{\mathcal{P}}(\mathbf{x})$ and $\mathbf{y}_n = F(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$. We then use the resulting GLLiM distribution denoted by p_G (and depending on $\boldsymbol{\theta}$) as a surrogate model for the pdf of $(\mathbf{X}, \mathbf{Y})$. Let's stress out that this first step does not use any prior information on $\mathbf{X}$.

The purpose is to exploit the tractable density p_G provided by the GLLiM model. Indeed, from p_G, conditional distributions are available in closed form and in particular:

$$p_G(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k^*(\mathbf{y}) \mathcal{N}(\mathbf{x}; \boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*) \tag{3}$$
$$\text{with } w_k^*(\mathbf{y}) = \frac{\pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j^* \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)}$$

where a new parametrization $\boldsymbol{\theta}^* = \{\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1:K}$ is used that can be easily deduced from $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}
\boldsymbol{c}_k^* &= \boldsymbol{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k \\
\boldsymbol{\Gamma}_k^* &= \boldsymbol{\Sigma}_k + \boldsymbol{A}_k \boldsymbol{\Gamma}_k \boldsymbol{A}_k^\top \\
\boldsymbol{\Sigma}_k^* &= \left(\boldsymbol{\Gamma}_k^{-1} + \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{A}_k\right)^{-1} \\
\boldsymbol{A}_k^* &= \boldsymbol{\Sigma}_k^* \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \\
\boldsymbol{b}_k^* &= \boldsymbol{\Sigma}_k^* \left(\boldsymbol{\Gamma}_k^{-1} \boldsymbol{c}_k - \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{b}_k\right)
\end{aligned} \tag{4}$$

The next section shows how to integrate the given prior information on $\mathbf{X}$.

2.2 Prediction step using prior information

We now observe that the target posterior (with prior information) can be factored into the product of the prior and a prior-less posterior:

$$p^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \mathcal{N}(\mathbf{x}; \mathbf{x}_0, \boldsymbol{\Gamma}_0)$$

where

$$p(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \propto \mathcal{U}_{\mathcal{P}}(\mathbf{x}) \mathcal{N}(F(\mathbf{x}); \mathbf{y}, \boldsymbol{\Sigma})$$

Since the GLLiM model has been learned such as to provide an approximation of $p(\mathbf{x}|\mathbf{Y} = \mathbf{y})$ through (3), we approximate the target posterior with

$$p_G^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{x}; \mathbf{x}_0, \boldsymbol{\Gamma}_0)\, p_G(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) \tag{5}$$

The key feature is that this density remains in closed form: it actually remains a Gaussian Mixture, with weights, means and covariances $(\alpha_k, \mathbf{x}_k, S_k)_{K=1\dots K}$ given by

$$\begin{aligned}
S_k &= \left(\boldsymbol{\Gamma}_0^{-1} + (\boldsymbol{\Sigma}_k^*)^{-1}\right)^{-1} \\
\mathbf{x}_k &= S_k \left(\boldsymbol{\Gamma}_0^{-1}\mathbf{x}_0 + (\boldsymbol{\Sigma}_k^*)^{-1} m_k^*\right) \\
\beta_k &= \sqrt{\frac{|S_k|}{(2\pi)^L |\boldsymbol{\Gamma}_0||\boldsymbol{\Sigma}_k^*|}} \exp\left(-\frac{1}{2}(m_k^* - \mathbf{x}_0)^\top B_k (m_k^* - \mathbf{x}_0)\right) \\
B_k &= (\boldsymbol{\Gamma}_0 + \boldsymbol{\Sigma}_k^*)^{-1} \\
\alpha_k &= \frac{w_k^*(\mathbf{y})\beta_k}{\sum_{k=1}^K w_k^*(\mathbf{y})\beta_k} \\
m_k^* &= \boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*
\end{aligned} \tag{6}$$

This means that, once the first learning step is done, inference on the posterior can be performed very efficiently: for example one can solve the assimilation problem by computing the mean of $p_G^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta})$, which is straightforward. An uncertainty estimation is also available by computing the variance.

Note that the same formulas can be recovered by observing that accounting for an initial guess $\mathbf{x}_0$ amounts to add in the observations $\mathbf{y}$ an additional observation $\mathbf{x}_0$. Then when using a Gaussian prior for $\mathbf{x}_0$, this combines well with the initial GLLiM model to lead to an *augmented* GLLiM model in dimension $L \times (L + D)$, defined by $(\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ being left unchanged and $(\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k)$ modified into

$$\tilde{\boldsymbol{A}}_k = \begin{pmatrix} \boldsymbol{A}_k \\ \mathbf{I}_L \end{pmatrix}, \tilde{\boldsymbol{b}}_k = \begin{pmatrix} \boldsymbol{b}_k \\ 0_L \end{pmatrix}, \tilde{\boldsymbol{\Sigma}}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k & \mathbf{0}_{D,L} \\ \mathbf{0}_{L,D} & \boldsymbol{\Gamma}_0 \end{pmatrix}$$

2.3 Extension to a more complex prior

So far, we have only considered a really simple prior distribution on $\mathbf{X}$. However, the result from the previous section can easily be extended to the case of Gaussian mixtures. Indeed, we can replace $\mathcal{N}(\mathbf{x}; \mathbf{x}_0, \boldsymbol{\Gamma}_0)$ by a Gaussian mixture with parameters $(a_i, \mu_i, \Gamma_i)_{i=1..I}$, and still obtain $p_G^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta})$ as a Gaussian mixture, this time with $K \times I$ components. Its parameters $(\alpha_{k,i}, \mathbf{x}_{k,i}, S_{k,i})$ are given by the following equations, which are a generalization of (6) :

$$
\begin{aligned}
S_{k,i} &= \left(\Gamma_i^{-1} + (\mathbf{\Sigma}_k^*)^{-1}\right)^{-1} \\
\mathbf{x}_{k,i} &= S_{k,i}\left(\Gamma_i^{-1}\mu_i + (\mathbf{\Sigma}_k^*)^{-1} m_k^*\right) \\
\beta_{k,i} &= \sqrt{\frac{|S_{k,i}|}{(2\pi)^L |\Gamma_i||\mathbf{\Sigma}_k^*|}} \exp\left(-\frac{1}{2}(m_k^* - \mu_i)^\top B_{k,i}(m_k^* - \mu_i)\right) \\
B_{k,i} &= (\Gamma_i + \mathbf{\Sigma}_k^*)^{-1} \\
\alpha_{k,i} &= \frac{w_k^*(\mathbf{y}) a_i \beta_{k,i}}{\sum_{k=1}^K \sum_{i=1}^I w_k^*(\mathbf{y}) a_i \beta_{k,i}} \\
m_k^* &= \boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*
\end{aligned}
\tag{7}
$$

This flexibility opens the door to more advanced inference tasks. However, in the following, we focus on the simpler case of a Gaussian prior, which is sufficient in the real world scenario we present in the next section.

3 Illustration on a detailed snowpack model

We present an application of our method to an example coming from [Gay et al., 2015], which study the snowpack composition through an electromagnetic backscattering model (EBM). More specifically, initial parameters values are coming from measurements performed manually by experts. The goal is then to refine these initial measurements using information available in the reflectivity measured by a radar. The quantities at stake relate to the composition of the snow layers, namely the snow diameter d_i and its density ρ_i for each layer $i = 1 : L$. Thus, given L layers, the parameters of interest $\mathbf{x}$ are of length $2 * L$. The backscattering measurement y is a scalar related to the parameters through $y = F(\mathbf{x}) = F_{EBM}(d_1, \ldots, d_L, \rho_1, \ldots, \rho_L)$. We refer to [Phan et al., 2014] for more details and the explicit expression of F_{EBM}.

Figure 1 shows the assimilation results for 4 snow carrots. The measurements come the from NoSREx report (measured at X-band, VV polarized, with an incidence angle of 40°). These results are only preliminary, but exhibit two properties that are consistent with previous findings. First, the same pattern for the diameters as in [Gay et al., 2015] is observed: the initial, expert measurements are consistently two high. Second, the density profiles, after assimilation, are increasing with the depth, which is physically sound.

4 Conclusion

We have proposed a Bayesian inversion approach to solve assimilation tasks. We have shown that the inverse regression approach GLLiM could be also adapted to account for a priori knowledge. This framework is especially interesting when we deal when the forward model is fixed, and assimilation is needed for a high number of observations, initial guesses

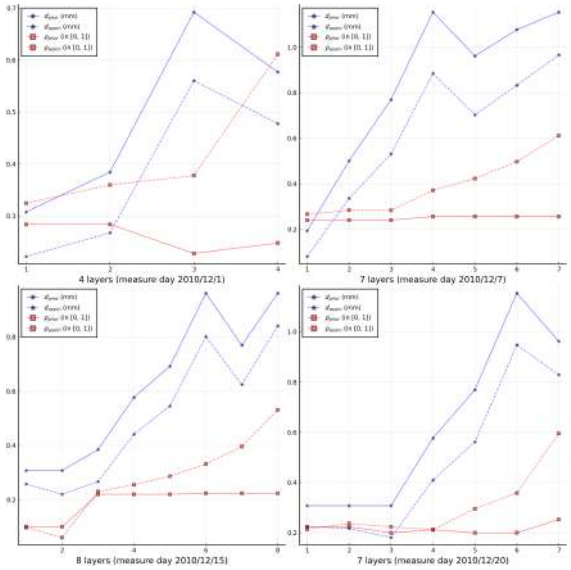


Figure 1: Snow layers properties assimilation. Layers depth is increasing from left to right (that is, the surface is on the left). Snow flakes diameter is in blue, density in red. Initial guesses are in solid line, assimilation result in dashed line.

or prior covariance levels, since the same learned GLLiM model can then be reused. In addition the possibility to use Gaussian mixtures as prior may cover a large range of physical constraints. Future work also includes the study of the covariance choice impact on the final assimilation results.

References

[Deleforge et al., 2015] Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911.

[Gay et al., 2015] Gay, M., Phan, X.-V., Ferro-Famil, L., Karbou, F., Durand, Y., Girard, A., and D 'urso, G. (2015). Simulation de la rétrodiffusion radar du manteau neigeux. Comparaison avec les données d'un radar SOL et TSX (projet NoSREx). In *L'Environnement ElectroMagnétique Des Radars à l'horizon 2020 : Quels Enjeux En Termes de Modélisation et Moyens de Mesures ?*, ENVIREM_FINAL, Gif sur Yvette, France.

[Phan et al., 2014] Phan, X. V., Ferro-Famil, L., Gay, M., Durand, Y., Dumont, M., Morin, S., Allain, S., D'Urso, G., and Girard, A. (2014). 1D-Var multilayer assimilation of X-band SAR data into a detailed snowpack model. *The Cryosphere*, 8(5):1975–1987.

PROJECTIONS D'INDICATEURS EPIDEMIOLOGIQUES A PARTIR D'UN MODELE MARKOVIEN DE TYPE ILLNESS-DEATH : APPLICATION A L'INFARCTUS DU MYOCARDE EN FRANCE JUSQU'EN 2035

Johann Kuhn[1] & Yann le Strat[2] & Christophe Bonaldi[3] & Clémence Grave[4] & Valérie Olié[5] & Pierre Joly[6]

[1] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – johann.kuhn@santepubliquefrance.fr*
[1] *Université Paris-Est Créteil, Créteil 94010, France*
[2] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – yann.lestrat@santepubliquefrance.fr*
[3] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – christophe.bonaldi@santepubliquefrance.fr*
[4] *Santé publique France, Direction des Maladies non Transmissibles et Traumatismes (DMNTT), Saint-Maurice 94415, France – clemence.grave@santepubliquefrance.fr*
[5] *Santé publique France, Direction des Maladies non Transmissibles et Traumatismes (DMNTT), Saint-Maurice 94415, France – valerie.olie@santepubliquefrance.fr*
[6] *Université de Bordeaux, ISPED, Centre Inserm I1219 – Bordeaux Population Health, Bordeaux 33076, France – pierre.joly@u-bordeaux.fr*

Résumé. Les maladies cardiovasculaires sont la principale cause de décès dans le monde et en Europe. En France, l'infarctus du myocarde est une cause importante de morbidité, de recours aux soins, d'altération de la qualité de vie et de mortalité. Une augmentation du nombre de cas d'infarctus du myocarde est attendue en France, engendrée par plusieurs phénomènes : le vieillissement de la population dans les décennies à venir, le maintien à un niveau élevé de la prévalence des facteurs de risque dans la population et la baisse de la mortalité cardiovasculaire. L'objectif du travail présenté est la projection, à partir des données du Système National des Données de Santé (SNDS), d'indicateurs épidémiologiques (incidence, prévalence) pour l'infarctus du myocarde jusqu'en 2035 en France. La méthodologie de projection se base sur un modèle markovien appelé « *illness-death* » utilisant des estimations d'incidence à partir d'un modèle âge-cohorte, des estimations de la mortalité et de taille de population dans la population générale à partir de données démographiques.

Mots-clés. Infarctus du myocarde, projection, incidence, prévalence, modèle markovien, modèle âge-période-cohorte

Abstract. Cardiovascular diseases are the main cause of death worldwide and in Europe. In France, myocardial infarction is a major cause of morbidity, recourse to healthcare, degradation of quality of life and mortality. An increase in the number of myocardial infarction cases is expected in France, caused by several phenomena: the ageing of the population in the coming decades, the continued high prevalence of risk factors in the population and the decline in cardiovascular mortality. The objective of the work presented is the projection, using data from the database "Système National des Données de Santé" (SNDS), of epidemiological indicators (incidence, prevalence) for myocardial infarction up to 2035 in France. The projection methodology is based on a Markov model called "illness-death" using incidence estimates from an age-cohort model, estimates of mortality and population size in the general population from demographic data.

Keywords. Myocardial infarction, projection, incidence, prevalence, Markov model, age-period-cohort model

1 Introduction

Les maladies cardiovasculaires sont la principale cause de décès dans le monde et en Europe. Selon le fardeau global des maladies (GBD) en 2019, la mortalité mondiale due aux maladies cardiovasculaires a été estimée à 18,6 millions de décès, soit 32,8 % de l'ensemble des décès dans le monde, contre 25,9 % en 1990 (Vos et al., 2020).

En France, les maladies cardiovasculaires représentent un fardeau extrêmement lourd puisqu'elles sont à l'origine annuellement de 140 000 décès (seconde cause de décès tous sexes confondus, mais première pour les femmes) et de 1 million de patients hospitalisés. De plus, elles exposent à des séquelles lourdes et invalidantes et impactent la qualité de vie des patients. Par ailleurs, les cardiopathies ischémiques dont l'infarctus du myocarde ont été identifiées comme la seconde cause d'années de vie perdues en France (Vos et al., 2020).

Une augmentation du nombre de cas d'infarctus du myocarde (IDM) est attendue en France en raison de plusieurs phénomènes. Le premier est le vieillissement de la population française qui entraîne une augmentation du nombre de personnes âgées de 65 ans et plus qui sont les principales victimes de la maladie (*Projections de population à l'horizon 2070 - Insee Première - 1619*, s. d.). Le deuxième est l'augmentation de l'incidence des hospitalisations pour IDM depuis le début des années 2000 chez les adultes de moins de 65 ans, avec une diminution de l'âge moyen lors de l'infarctus du myocarde (Amélie Gabet et al., 2017). Cette augmentation de l'incidence devrait se poursuivre compte tenu du maintien à un niveau élevé des principaux facteurs de risque (cholestérol, hypertension, tabac, diabète, obésité) dans la population (Blacher et al., 2020; Lailler et al., 2020; Pasquereau A, 2020; Vallée et al., 2020; Verdot et al., 2017). Enfin, la baisse de mortalité par infarctus du myocarde depuis plusieurs décennies concoure à l'augmentation de la prévalence de l'infarctus du myocarde dans la population.

Dans ce contexte, la projection du nombre de cas de l'infarctus du myocarde devient un enjeu important de santé publique en France. L'augmentation importante de la prévalence de cette pathologie dans les années à venir nécessite d'être anticipée afin de permettre une prise en charge des patients. Ceux-ci porteront sur la prise en charge hospitalière et thérapeutique : en phase aigüe, l'infarctus du myocarde entraîne des hospitalisations longues et une prise en charge en soins de suite et réadaptation pour la réadaptation cardiaque. Les patients ayant eu un infarctus du myocarde, peuvent avoir une dégradation de leur qualité de vie jusque dans leurs activités quotidiennes (De Peretti et al., 2014), cette pathologie pouvant entraîner les patients vers un état de dépendance. Ainsi, l'anticipation des besoins en termes d'accueil de ces personnes et d'accompagnement est également nécessaire.

Dans ce travail, nous projetons plusieurs indicateurs épidémiologiques (nombre de cas prévalents hospitalisés, prévalence, âge moyen des cas incidents) jusqu'en 2035. Nous nous sommes basés sur une méthodologie de projection de la prévalence développée pour les maladies chroniques en utilisant un modèle "*illness-death*" multi-états appliqué aux données d'hospitalisation issues du Système National des Données de Santé (SNDS) (Joly et al., 2013). Nous allons d'abord décrire les données SNDS utilisées pour les projections. Une description du modèle, le choix des hypothèses et la méthode d'estimation seront expliqués dans la section 3. Enfin, nous terminerons par une discussion sur l'objectif de ce travail et la pertinence du modèle.

2 Données

Entre 2007 et 2015 en France métropolitaine, 519 490 patients (67.4 % pour les hommes, 32.6 % pour les femmes) âgés de 35 à 95 ans, ont été hospitalisés pour un infarctus du myocarde (IDM). Ces patients ont été identifiés dans la base du PMSI (Programme de Médicalisation des Systèmes

d'Information). Les patients étaient sélectionnés lorsqu'ils avaient eu une hospitalisation complète (au moins une nuit d'hôpital) avec un diagnostic principal d'IDM (I21-I23) selon la classification internationale des maladies, dixième version (CIM-10), et n'ayant pas eu d'antécédents médicaux d'IDM au cours des deux années précédentes selon l'année d'inclusion. Le taux annuel brut d'incidence des hospitalisations par âge et par sexe a été calculé en utilisant les estimations annuelles des populations moyennes nationales de l'Institut national de la statistique et des études économiques (INSEE). Les données démographiques (taux de mortalité, taille de la population) entre 1955 et 2070 par âge (35 à 95 ans) et par sexe sont issues de l'INSEE.

3 Modèle

La méthode de projection se base sur un modèle markovien à trois états, appelé "*illness-death*" (Figure 1), qui décrit les transitions possibles entre trois états : état 0 (vivant, sans IDM), état 1 (malade (IDM)), état 2 (décédé). Le modèle est dit irréversible car un individu ne peut pas guérir de sa maladie. Les transitions sont caractérisées par leur intensité de transition que nous notons α_{01} pour le taux d'incidence de l'infarctus du myocarde, α_{02} pour le taux de mortalité des non-malades et α_{12} pour le taux de mortalité des malades. De manière générale, les intensités de transition dépendent du temps calendaire *t*.

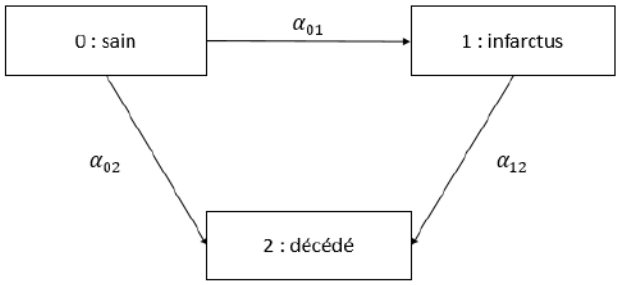


Figure 1 : Modèle « illness-death »

3.1 Hypothèses

Nous avons choisi l'hypothèse d'une non-homogénéité du temps pour les intensités de transition. La non-homogénéité du temps indique que les intensités de transition ne sont pas constantes au cours du temps calendaire.

3.2 Estimation des intensités de transition

Le taux d'incidence des individus hospitalisés pour un IDM, noté $\alpha_{01}(a,c)$ pour un âge a et une cohorte c (appelée aussi génération), a été estimé à partir d'un modèle âge-cohorte :

$$Y(a,c) \sim Poisson\big(m(a,c) \times \alpha_{01}(a,c)\big)$$

$$\log(\alpha_{01}(a,c)) = \alpha_a + \gamma_c$$

où $Y(a,c)$ est le nombre de cas incidents de l'IDM, $m(a,c)$ est le nombre de personnes-années, α_al'effet de l'âge et γ_c l'effet cohorte. Ce modèle a été comparé avec d'autres modèles, notamment le modèle âge-période. Le choix du modèle s'est basé sur le critère d'information d'Akaike (AIC), l'adéquation des données prédites aux données observées et la forme des projections calculées par le modèle. Le taux d'incidence a été estimé à l'aide d'un modèle linéaire généralisé (GLM) incluant des splines naturelles. De plus, nous avons considéré que le taux d'incidence était nul avant un âge a_0 (35 ans ici) en se basant sur l'épidémiologie de l'IDM :

$$\alpha_{01}(a,c) = 0 \text{ for } a \leq a_0$$

Nous avons considéré que la mortalité des individus sains était équivalente à la mortalité dans la population générale obtenue à partir des données de l'INSEE que l'on dénote α_2. Cette hypothèse est plausible car la prévalence de l'IDM n'est pas élevée (inférieure à 10 %). Comme l'âge, le temps calendaire et la cohorte sont liés par la relation suivante : $t = a + c$, il est alors possible d'obtenir le taux de mortalité des individus sains dépendant de l'âge et de la cohorte à partir du taux de mortalité dépendant de l'âge et du temps calendaire. Le taux de mortalité des individus sains par âge et par cohorte peut alors s'écrire :

$$\alpha_{02}(a,c) = \alpha_2(a,c)$$

Nous avons ensuite ajusté un modèle de Gompertz-Makeham aux taux de mortalité de l'INSEE pour obtenir une fonction continue de l'âge pour chaque cohorte. (Olshansky & Carnes, 1997)

La mortalité des individus malades a été considérée comme proportionnelle à la mortalité des individus sains avec un risque relatif dépendant à la fois de l'âge et de la durée depuis l'apparition de la maladie. Notons RR_{d_1} le risque relatif associé à la première année suivant l'occurrence de la maladie et RR_{d_2} le risque relatif par la suite. La mortalité d'un individu ayant eu un infarctus jusqu'à l'année suivante s'écrit :

$$\alpha_{12}(a,c) = \alpha_{02}(a,c) \times RR_{d_1}$$

et pour un individu ayant eu un infarctus et au-delà de l'année suivant l'occurrence :

$$\alpha_{12}(a,c) = \alpha_{02}(a,c) \times RR_{d_2}$$

Nous avons identifié les risques relatifs à partir d'une cohorte danoise dans laquelle 3.092.580 personnes non diabétiques âgées d'au moins 30 ans et sans antécédents médicaux d'infarctus du myocarde ont été incluses et suivies entre 1997 et 2006. (Norgaard et al., 2010)

Le nombre de cas prévalents pour une année t est le nombre de personnes malades de tous âges à l'année t considérée. Notons a_0 l'âge minimal et a_{max} l'âge maximal des individus (respectivement égaux à 35 et 95 ans pour ce travail), $\nu(a_0, t)$ la taille de la population pour un âge a_0 et une année t et c la cohorte, la prévalence s'écrit alors :

$$N_{prev}(t) = \sum_{k=a_0}^{a_{max}} \nu(a_0, t - a_{max} + k) \times P_{01}(a_0, a_{max} - k + a_0 | c)$$

avec $P_{01}(a_0, a_{max}|c)$ la probabilité de se trouver dans l'état 1 (infarctus) en $a_{max} - k + a_0$ sachant que l'individu était dans l'état 0 (sain) en a_0.selon la cohorte.

La probabilité $P_{01}(a_0, a_{max}|c)$ est définie par :

$$P_{01}(a_0, a_{\max}|c) = \int_{a_0}^{a_{max}} e^{-A_{01}(a_0,u,c) - A_{02}(a_0,u,c)} \; a_{01}(u,c) \; e^{-A_{12}(u,a_{max},c)} \; du$$

avec A_{01}, A_{02}, A_{12} les intensités de transition cumulées pour les transitions de l'état 0 vers l'état 1,

de l'état 0 vers l'état 2 et de l'état 1 vers l'état 2, respectivement. Pour alléger les formules, nous ne définissons que la formule générale de l'intensité de transition cumulée :

$$A_{kl}(s_1, s_2) = \int_{s_1}^{s_2} \alpha_{kl}(u)du$$

où s_1, s_2 deux temps tels que $s_1 < s_2$.

L'âge moyen des cas incidents est définie par :

$$\overline{age_{(inci)}}(t) = \frac{\sum_{a_0}^{a_{max}} a \times P_{00}(a_0, a|c) \times \alpha_{01}(a, c) \times \nu(a_0, c)}{\sum_{a_0}^{a_{max}} P_{00}(a_0, a|c) \times \alpha_{01}(a, c) \times \nu(a_0, c)}$$

avec $\nu(a_0, c)$ la population d'âge a_0 pour une cohorte c et $P_{00}(a_0, a|c)$ la probabilité de rester dans l'état 0 (sain) entre les âges a_0 et a selon la cohorte des individus.

De même, $P_{00}(a_0, a|c)$ est définie par :

$$P_{00}(a_0, a|c) = e^{-A_{01}(a_0,a,c) - A_{02}(a_0,a,c)}$$

4 Discussion

Les modèles de la famille âge-période-cohorte comme le modèle âge-cohorte sont fréquemment utilisés pour modéliser des taux d'incidence et de mortalité dans la littérature (Carstensen, 2007; Clayton & Schifflers, 1987) et permettent d'exprimer le taux d'incidence en fonction de l'âge et d'un effet du temps comme la cohorte. Malgré la période de suivi des patients hospitalisés assez courte (2007 à 2015 soit 9 ans), nous obtenons des taux d'incidence du même ordre de grandeur que l'on peut retrouver dans la littérature (*L'état de santé de la population en France - RAPPORT 2017 - Ministère des Solidarités et de la Santé*, s. d.). Pour ces projections, nous avons sélectionné le modèle simple qu'est l'âge-cohorte car il nous permettait d'avoir le meilleur compromis entre l'adéquation des données prédites aux données observées et l'allure des projections parmi différents modèles testés de la même famille.

De nombreuses hypothèses du modèle ont été fixées avant de calculer les estimations des différents indicateurs épidémiologiques. Il est important d'évaluer *a posteriori* si nos hypothèses choisies pour le modèle n'ont pas été trop restrictives et fortes. Les estimations des différents indicateurs (prévalence et âge moyen des cas incidents) valident nos hypothèses puisque nous retrouvons des valeurs du même ordre de grandeur dans la littérature. (A Gabet et al., 2016; *L'état de santé de la population en France - RAPPORT 2017 - Ministère des Solidarités et de la Santé*, s. d.). De ce fait, ce travail pourrait s'appliquer à d'autres maladies chroniques cardiovasculaires comme les accidents vasculaires cérébraux.

Références

Blacher, J., Gabet, A., Vallée, A., Ferrières, J., Bruckert, E., Farnier, M., & Olié, V. (2020). Prevalence and management of hypercholesterolemia in France, the Esteban observational study. *Medicine*, *99*(50), e23445. https://doi.org/10.1097/MD.0000000000023445

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in medicine*, *26*(15), 3018-3045.

Clayton, D., & Schifflers, E. (1987). Models for temporal variation in cancer rates. I: age–period and age–cohort models. *Statistics in medicine*, *6*(4), 449-467.

De Peretti, C., DANCHIN, N., DANET, S., OLIE, V., & Gabet, A. (2014). Prévalences et statut fonctionnel des cardiopathies ischémiques et de l'insuffisance cardiaque dans la population adulte en France : Apports des enquêtes déclaratives Handicap-Santé. *Bulletin épidémiologique hebdomadaire*, *9-10*, 172-181.

Gabet, A, Danchin, N., & Olié, V. (2016). Infarctus du myocarde chez la femme : Évolutions des taux d'hospitalisation et de mortalité, France, 2002-2013. *Bull Épidémiologique Hebd*, *8*, 7-8.

Gabet, Amélie, Danchin, N., Juillière, Y., & Olié, V. (2017). Acute coronary syndrome in women : Rising hospitalizations in middle-aged French women, 2004–14. *European Heart Journal*, *38*(14), 1060-1065.

Joly, P., Touraine, C., Georget, A., Dartigues, J.-F., Commenges, D., & Jacqmin-Gadda, H. (2013). Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics*, *69*(1), 109-117.

Lailler, G., Piffaretti, C., Fuentes, S., Nabe, H. D., Oleko, A., Cosson, E., & Fosse-Edorh, S. (2020). Prevalence of prediabetes and undiagnosed type 2 diabetes in France : Results from the national survey ESTEBAN, 2014-2016. *Diabetes Research and Clinical Practice*, *165*, 108252. https://doi.org/10.1016/j.diabres.2020.108252

L'état de santé de la population en France—RAPPORT 2017—Ministère des Solidarités et de la

Santé. (s. d.). Consulté 20 décembre 2020, à l'adresse https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/recueils-ouvrages-et-rapports/recueils-annuels/l-etat-de-sante-de-la-population/article/l-etat-de-sante-de-la-population-en-france-rapport-2017

Norgaard, M. L., Andersen, S. S., Schramm, T., Folke, F., Jørgensen, C., Hansen, M. L., Andersson, C., Bretler, D., Vaag, A., Køber, L., & others. (2010). Changes in short-and long-term cardiovascular risk of incident diabetes and incident myocardial infarction—A nationwide study. *Diabetologia*, *53*(8), 1612-1619.

Olshansky, S. J., & Carnes, B. A. (1997). Ever since gompertz. *Demography*, *34*(1), 1-15.

Pasquereau A, N.-T. V., Andler R, Arwidson P, Guignard R. (2020). Consommation de tabac parmi les adultes : Bilan de cinq années de programme national contre le tabagisme, 2014-2019. *Bull Epidémiol Hebd. 2020*, *14*, 273-281.

Projections de population à l'horizon 2070—Insee Première—1619. (s. d.). Consulté 20 décembre 2020, à l'adresse https://www.insee.fr/fr/statistiques/2496228

Vallée, A., Gabet, A., Grave, C., Sorbets, E., Blacher, J., & Olié, V. (2020). Patterns of hypertension management in France in 2015 : The ESTEBAN survey. *Journal of Clinical Hypertension (Greenwich, Conn.)*, *22*(4), 663-672. https://doi.org/10.1111/jch.13834

Verdot, C., Torres, M., Salanave, B., & Deschamps, V. (2017). Corpulence des enfants et des adultes en France métropolitaine en 2015. Résultats de l'étude Esteban et évolution depuis 2006. *Bull Epidémiol Hebd*, *13*, 234-241.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H., Aboyans, V., Abrams, E. M., Abreu, L. G., Abrigo, M. R. M., Abu-Raddad, L. J., Abushouk, A. I., … Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019 : A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1204-1222.

https://doi.org/10.1016/S0140-6736(20)30925-9

Meta-modélisation multi-fidélité avec des séries-temporelles en sortie

Baptiste Kerleguer[1,2] & Claire Cannamela [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France baptiste.kerleguer@cea.fr* [2] *Centre de Mathématiques Appliquées, Ecole polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France*

Résumé. Nous examinons la meta-modelisation d'un code numérique complexe dans un cadre multi-fidélité lorsque la sortie du code est une série temporelle. En utilisant les données basse et haute fidélité, nous proposons une méthode originale de régression par processus gaussien. La sortie du code est développée sur une base construite à partir des données. Les premiers coefficients de l'expansion de la sortie du code sont traités par une approche de co-krigeage. Les derniers coefficients sont traités collectivement par une approche de krigeage avec tensorisation de covariance. Le meta-modèle qui en résulte, prenant en compte l'incertitude dans la construction de la base, s'avère plus performant en termes d'erreurs de prédiction et de quantification de l'incertitude que les techniques standard de réduction des dimensions.

Mots-clés. Processus gaussiens, sorties séries-temporelles, covariance tensorisée, réduction des dimensions

Abstract. we consider the surrogate modeling of a complex numerical code in a multi-fidelity framework when the code output is a time series. Using low- and high-fidelity data, an original Gaussian process regression method is proposed. The code output is expanded on a basis built from the experimental design. The first coefficients of the expansion of the code output are processed by a co-kriging approach. The last coefficients are collectively processed by a kriging approach with covariance tensorization. The resulting surrogate model taking into account the uncertainty in the basis construction is shown to have better performance in terms of prediction errors and uncertainty quantification than standard dimension reduction techniques.

Keywords. Gaussian processes, time-series outputs, tensorized covariance, dimension reduction

1 Introduction

Les progrès de la modélisation scientifique ont conduit au développement de codes plus complexes et plus coûteux en termes de calcul. Il est donc devenu nécessaire d'utiliser des méta-modèles, construits à partir des sorties de ces codes, afin d'étudier les incertitudes de

ces codes. Nous souhaitons aborder le cas d'un code complexe avec des séries temporelles en sorties. De plus il existe de plus en plus de codes modélisant le même système ce qui a permit d'introduire la multi-fidélité. Nous étudions de cas de la multi-fidélité hiérarchisée où les codes peuvent être classé en fonction de leur coût et de leur précision.

Une méthode très employée dans la communauté de la quantification d'incertitudes pour la réalisation de méta-modèles est la régression par processus gaussien. Cette méthode, également appelé krigeage, a été introduite pour la géostatistique avant d'être utilisé pour les expériences numériques et en quantification d'incertitudes, Avec l'émergence de la multi-fidélité il est devenu intéressant de construire des approches multi-fidélité pour la construction de modèles de substitution. Le schéma auto-régressif présenté par Kennedy et O'Hagan (2000) est le premier résultat important. Ensuite Le Gratiet (2013) à modifier cette méthode pour permettre au co-krigeage multi-fidélité d'être décomposable en krigeage indépendants.

Parmi les codes ayant des sorties de grande dimension nous nous intéressons à ceux dont les sorties sont des fonctions dépendant du temps. Lorsqu'elles sont échantillonnées elles sont appelées séries-temporelles. On fait la supposition que l'échantillonnage est fait sur une grille fine et régulière donc la sortie est de très grande dimension. Les séries temporelles sont traités en utilisant des méthodes simple fidélité.

Pour construire des méta-modèles à sortie série-temporelle deux méthodes ont étés envisagées, soit réduire la dimension de la sortie ou adapter le noyau de régression, voir Perrin (2020). Pour la réduction il est plus difficile de quantifier l'incertitude en particulier quand l'ensemble d'arrivée du code est mal connu. De plus une grande quantité de données basse fidélité peut rendre le temps de calcul très long sans améliorer sensiblement le résultat. Le noyau pose le problème de la séparation de la variable temporelle et des variables d'entré, ce qui réduit les cas d'application. Ainsi nous proposons de combiner la réduction de dimensions sur une partie de le sortie du code et pour le reste on suppose qu'elle suit un processus gaussien de covariance tensoriée.

2 Multi-fidélité pour séries temporelles

On suppose que les code basse et haute fidélité sont la réalisation d'un processus stochastique (Z_L, Z_H). Ils sont connues pour les valeurs x_L et x_H sur la grille t de N_t points. On suppose que l'on choisit une famille libre de N fonction. On construit donc les processus stochastique suivant:

$$Z_L(x,t) = \sum_{i=1}^{N} A_{L,i}(x)\Gamma_i(t) + Z_L^{\perp}(x,t), \tag{1}$$

$$Z_H(x,t) = \sum_{i=1}^{N} A_{H,i}(x)\Gamma_i(t) + Z_H^{\perp}(x,t), \tag{2}$$

avec $(A_{L,i}, A_{H,i})$ un processus gaussien que nous modélisons par un modèle auto-régressif. Les γ_i sont éléments de la base temporelles de l'ensemble S_N. Notre étude a considéré alternativement que la base était stochastique ou déterministe. $Z_H^{\perp}(x,t)$ est la partie orthogonale que l'on ignore en basse fidélité et qui est supposé être un processus gaussien à covariance tensorisée. Pour prédire $Z_H(x,t)$ nous avons donc 3 étapes:

- La définition des fonctions Γ_i. Pour cela nous avons proposé deux méthodes se fondant sur la décomposition en valeurs singulières. La première possibilité est de réaliser une décomposition des données basse fidélité et de prendre les N premiers vecteurs propres. La deuxième méthode est similaire à la validation croisée et consiste à réaliser la moyenne et la variance de chacune des bases construites avec la méthode précédente mais sur uniquement une partie des éléments basse fidélité.
- Régression par processus gaussien par le modèle AR(1) à $(A_{L,i}, A_{H,i})$. Pour cela nous utilisons la méthode décrite par Le Gratier (2013). Ainsi N modèles AR(1) décorrélés sont construit indépendamment.
- La régression sur la partie orthogonale $Z_H^{\perp}(x,t)$. Pour cela nous utilisons la méthode décrite par Perrin (2020). On a montré que la prédiction et la moyenne de prédiction reste sur l'espace $S_N^{\perp} = \text{span}\{\Gamma_1, \ldots, \Gamma_N\}$ and $S_N^{\perp} = \text{span}\{\Gamma_{N+1}, \ldots, \Gamma_{N_t}\}$, espace orthogonale à S_N.

Pour le choix de N on peut calculer pour les premières valeurs et ainsi prendre la valeur qui donne le meilleur modèle sur le set d'apprentissage. En effet, le calcul d'une nouvelle valeur de N ne nécessite que le calcul d'un co-krigeage et d'un krigeage avec covariance tensorisée. De plus pour des N plus grand que la taille de l'ensemble haute fidélité on peut montrer que les performances seront de moins en moins bonnes donc on sait que $N \leq N_H$.

3 Résultat

Notre méthode est testée sur un code physique simple d'un pendule liée à un système masse ressort. On fait la mesure de la position du pendule en fonction du temps qui est notre sortie. On dispose de $N_L = 100$ et $N_H = 10$ expériences numériques basse et haute fidélité. On dispose d'une grille régulière de $N_t = 101$ échantillons sur le segment $[0, 10]$ qui correspond au problème physique. Pour évaluer la pertinence de notre modèle nous avons choisi de comparé les valeurs de $Q^2(t)$. Cet indicateur est défini comme:

$$Q^2(t) = 1 - \frac{\sum_{i=1}^{N_H}(z_H(x_i,t) - \widehat{z_H(x_i,t)})^2}{N_H \,\text{Var}(z_y(x,t))}, \tag{3}$$

avec x_i la i-ème valeur de l'ensemble de test, $z_H(x_i,t)$ la valeur sortie de code haute fidélité au point x_i, $\widehat{z_H(x_i,t)}$ est la prédiction du modèle à tester pour le point x_i et Var représente la variance sur l'ensemble des points connus.

On peut conclure dans notre exemple que la multi-fidélité est nécessaire pour car le modèle simple fidélité présentée à la figure 1 montre des résultats insuffisants. La méthode empirique de réduction de la dimension, également appelée par projection, est moins performante que les deux méthodes que nous mettons en avant. Les deux méthodes sont fondées sur notre méthode mais utilisent de manière de construire la base de fonctions temporelles différentes. La méthode Dirac utilise l'ensemble des données basse fidélité pour définir les Γ_i qui sont donc déterministes et la méthode Empirique construit des sous-ensembles pour avoir estimé l'incertitude sur les Γ_i que l'on suppose être les colonnes d'une matrice orthogonale aléatoire. Les noyaux de covariance du modèle AR(1) sont tous des Matèrns 5/2. Pour la covariance de la partie orthogonale la partie dépendant de x a un noyau de Matèrn 5/2 alors que la partie définie en t est calculé par maximum de vraisemblance à partir des données et de la covariance en x.

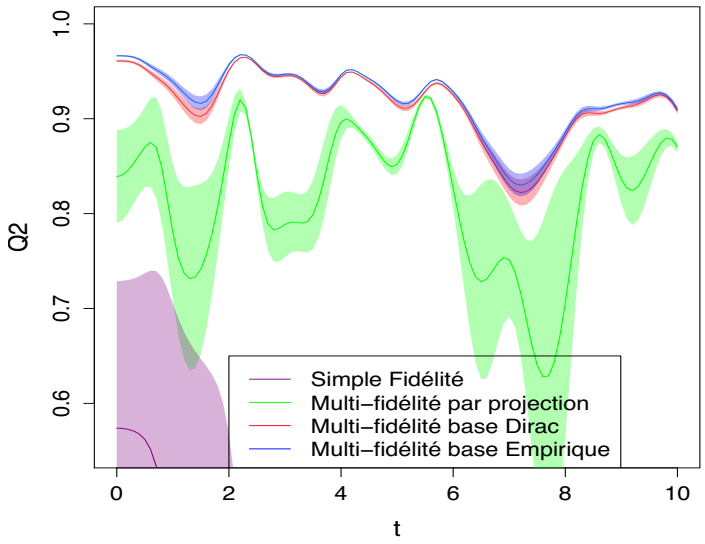


Figure 1: Comparaison des Q^2 en fonction du temps. La courbe foncée représente la moyenne de 40 différentes prédictions sur les données d'apprentissage différentes évaluées sur un même ensemble de test de 1000 expériences. La surface colorée représente 1.96 écart-type du Q^2 en dépendant du temps.

La méthode proposée pour la méta-modélisation multi-fidélité prend en compte toutes les incertitudes et prédit de manière très efficace les sorties du code de calcul pour des séries temporelles en sortie. Par rapport aux méthodes de la littérature elle permet de résoudre un problème nouveau avec de meilleurs garantis théoriques et a de meilleurs résultats sur un exemple.

Bibliographie

M. Kennedy et A. O'Hagan, (2000), Predicting the output from a complex computer code when fast approximations are available, Biometrika.
L. Le Gratiet, (2013), Bayesian analysis of hierarchical multifidelity codes, SIAM/ASA Journal on Uncertainty Quantification.
G. Perrin, (2020) Adaptive calibration of a computer code with time-series output, Reliability Engineering and System Safety.

Meta-modélisation multi-fidélité combinant processus Gaussiens et réseau de neurones bayésien

Baptiste Kerleguer[1,2] & Claire Cannamela [1] & Josselin Garnier [2]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France baptiste.kerleguer@cea.fr* [2] *Centre de Mathématiques Appliquées, Ecole polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France*

Résumé. Le problème d'intérêt est la méta-modélisation de la réponse d'un code informatique dans un cadre multi-fidélité, c'est-à-dire lorsque la sortie peut être évaluée à différents niveaux de prédiction et de temps de calcul. En utilisant des niveaux de code de haute et basse fidélité, nous avons développé une méthode originale combinant la régression par processus gaussien et le réseau neuronal bayésien. Le méta-modèle qui en résulte a des capacités de prédictions très satisfaisantes mais surtout il tient compte des incertitudes du processus gaussien et du réseau de neurones Bayésien.

Mots-clés. Processus Gaussiens, Réseau de neurones bayésien

Abstract. We want to build a surrogate model of a computer code in a multi-fidelity framework, i.e. when the output can be evaluated at different levels of prediction and calculation time. Using high and low fidelity code levels, we have developed an original method combining Gaussian process regression and Bayesian neural network. The resulting surrogate model has good prediction capabilities, but more importantly it takes into account the uncertainties of the Gaussian process and the Bayesian neural network.

Keywords. Gaussian process, Bayesien Neural Network

1 Introduction

Il existe de nombreux exemples d'utilisation de régression par processus gaussien dans un cadre multi-fidélité dérivée de l'article de Kennedy et O'Hagan (2000). Dans cet article une relation linéaire est supposée entre la sortie du code haute et celle de basse fidélité. Par la suite des relations plus complexes entre codes ont été introduites, par exemple par Perdikaris et *al* (2017). Ces méthodes, basées sur la régression par processus gaussien, permettent de quantifier les incertitudes de prédiction. Elles sont également très efficaces dans des contextes de données peu nombreuses, très fréquent lorsque le code est très coûteux en temps de calcul. Leur principale limite sont le besoin d'hypothèses sur la régularité de la relation entre les sorties de différents niveaux de fidélité et surtout la très grande difficulté de passer en grande dimension (en entrée et en sortie du modèle).

En parallèle le développement des réseaux de neurones a permis de construire des méta-modèles très performants et notamment des méta-modèles multi-fidélité. Les méta-modèles sont bien plus complexes à paramétrer mais permettent de traiter des problèmes de très grande dimension en entrée et en sortie. Or des avancées récentes notamment sur les réseaux de neurones bayésiens ont permis de quantifier les incertitudes de méta-modélisation, Meng et *al* (2020). Ainsi on peut espérer construire des modèles qui permettent la quantification des incertitudes dans un cadre de la grande dimension d'entrée mais aussi de sortie.

2 Méthode

Le méta-modèle est construit en deux parties. Dans un premier temps c'est seulement sur la sortie du code basse fidélité que nous allons construire un méta-modèle. Pour cela nous avons utilisé la régression par processus gaussiens. Puis un deuxième modèle a pour entrées les entrées du code et les sorties du code basse fidélité. La prédiction de la sortie haute fidélité fera ainsi appel aux deux méta-modèles.

Pour interfacer les deux modèles nous devons transmettre le premier méta-modèle au second méta-modèle, avec une prise en compte, pour ce dernier des résultats du premier au moment de la phase d'apprentissage. La première méthode envisagée a été d'utiliser la moyenne et la variance de sortie du modèle basse fidélité mais cela ne garantira pas le suivi des incertitudes. La seconde méthode envisagée et finalement adoptée est de s'intéresser à des réalisations de la loi de la sortie basse fidélité. Cette loi étant connue (sachant les données), la quadrature de Gauss-Hermite est très adaptée pour minimiser le nombre de réalisations nécessaires et de cette façon optimiser le temps de calcul. En ce qui concerne la régression par processus gaussien, nous avons opté pour un noyau de Matèrn, même s'il n'y a aucune limite au choix du noyau. Le réseau de neurones bayésien que nous utilisons est activé par une fonction ReLU avec une seule couche cachée afin de réduire au maximum les hyper-paramètres car nous avons peu de données. Nous nous sommes donnés des a priori classiques dans la littérature. Les sorties du réseau de neurones bayésien sont pondérées en fonction des points donnés par la formule de quadrature.

3 Résultats

Nous présentons des résultats pour des codes à deux niveaux de fidélité ainsi que des sorties scalaires. Les codes que nous utilisons sont des fonctions analytiques dont la dimension d'entrée varie de 1 à 4. Nous nous plaçons dans un contexte où les hypothèses de linéarité entre les codes ne sont pas validées et les résultats de ces modèles sont très mauvais. Ainsi nous devons aussi nous comparer à des modèles plus complexes.

Les performances de notre approche semblent comparables à celles obtenues à l'aide de réseaux de neurones. Ils présentent l'avantage de pouvoir quantifier les incertitudes en

fournissant une méthode d'échantillonnage de la loi à postériori. Nous espérons pouvoir monter en dimension de sortie et d'entrées mais le nombre d'hyper-paramètres nous limite encore dans un contexte de données peu nombreuses.

Bibliographie

M. Kennedy et A. O'Hagan, (2000), Predicting the output from a complex computer code when fast approximations are available, *Biometrika*.
P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence et G. E. Karniadakis, (2017), Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*.
X. Meng, H. Babaee et G. E. Karniadakis, (2020), Multi-fidelity Bayesian Neural Networks: ALgorithms and Applications, *arXiv preprint arXiv:2012.13294*.

Differentiation implicite pour la calibration de modèles non lisses

Quentin Klopfenstein [1] & Quentin Bertrand [2] & Mathieu Blondel [3] &
Samuel Vaiter [1,4] & Alexandre Gramfort [2] & Joseph Salmon [5]

[1] *IMB, Université de Bourgogne Franche-Comté, Dijon, France*
[2] *Université Paris-Saclay, Inria, CEA, Palaiseau, France*
[3] *Google research, Brain team, Paris, France*
[4] *CNRS*
[5] *IMAG, Université de Montpellier, CNRS, Montpellier, France*

Résumé. Trouver la valeur optimale d'hyperparamètres pour un modèle peut être écrit comme un problème d'optimisation à deux niveaux. Ce problème d'optimisation est très souvent résolu en utilisant des techniques de *grid-search*, *random-search* ou de l'optimisation bayésienne. Toutes ces méthodes peuvent être vues comme de l'optimisation à l'ordre zéro (sans gradient) mais sont difficilement utilisables lorsque le nombre d'hyperparamètres à sélectionner devient grand. Des méthodes d'optimisation du premier ordre peuvent surmonter ces difficultés : l'étape clés étant le calcul d'hypergradients, *i.e.* de gradients en fonction des hyperparamètres. Ces méthodes ont été très étudiées pour des modèles basés sur des problèmes d'optimisation lisses, cependant la littérature concernant les problèmes d'optimisation non lisses est plus rare. Dans ce travail, nous étudions les techniques classiques de calcul d'hypergradient (différenciation forward et implicite) lorsque le problème d'optimisation sous-jacent est convexe mais non lisse. Les résultats sur des modèles de régression et de classification montrent des gains significatifs en rapidité de calcul, en particulier lorsque le nombre d'hyperparamètres est grand.

Mots-clés. Optimisation d'hyperparamètres, Optimisation non-lisse, Hypergradients

Abstract. Finding the optimal hyperparameters of a model (a.k.a. model selection) can be cast as a bilevel optimization problem, which is typically solved using grid-search, random-search or Bayesian optimization. These methods can be seen as zero-order (*i.e.* gradient-free) techniques, and scale poorly with the number of hyperparameters to tune. First-order optimization methods can overcome these limitations, the key step being the computation of *hypergradients*, *i.e.* gradients *w.r.t* to the hyperparameters. Such methods have been largely studied for models based on smooth optimization problems, however the literature regarding non-smooth optimization problems is scarcer. In this work we study classical hypergradient computation techniques (implicit and iterative differentiation) when the underlying optimization problem is convex but non-smooth. Results on regression and classification problems reveal clear computational benefits, especially when multiple hyperparameters are required.

Keywords. Hyperparameter optimization, Non-smooth optimization, Hypergradients

1 Introduction

En apprentissage automatique, presque tous les modèles dépendent d'au moins un hyperparamètre. Le choix de ce dernier impacte fortement la qualité du modèle et sa performance. C'est le cas pour beaucoup d'estimateurs utilisés en apprentissage automatique, où un paramètre de régularisation permet de contrôler le poids de la régularisation par rapport au terme de fidélité aux données. Parmi ces estimateurs, on trouve la régression Ridge [3], le Lasso [10], l'elastic net [12], la regression logisitque pénalisée ℓ_1 [7] et les algorithmes de vecteurs supports [1]. Tous ces estimateurs sont solutions de problèmes d'optimisation qui peuvent s'écrire de manière général sous la forme:

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \Phi(\beta, \lambda) \triangleq f(X\beta) + \sum_{j=1}^{p} g_j(\beta_j, \lambda) \ , \tag{1}$$

avec $X \in \mathbb{R}^{n \times p}$ la matrice de *design*, $f : \mathbb{R}^n \to \mathbb{R}$ une fonction à gradient Lipschitz, $g_j(\cdot, \lambda)$ des fonctions convexes (possiblement non-lisse) et un paramètre de régularisation (ou hyperparamètre) $\lambda \in \mathbb{R}^r$. Une méthode classique de sélection d'hyperparamètre est de faire de l'optimisation d'hyperparamètre, c'est-à-dire choisir le paramètre λ de sorte que $\hat{\beta}^{(\lambda)}$ minimise un critère donné $\mathcal{C} : \mathbb{R}^p \to \mathbb{R}$ mesurant la performance du modèle. Plus formellement l'optimisation d'hyperparamètre peut s'écrire sous la forme d'un problème d'optimisation à deux niveaux.

$$\begin{aligned} &\arg\min_{\lambda \in \mathbb{R}^r} \{\mathcal{L}(\lambda) \triangleq \mathcal{C}\left(\hat{\beta}^{(\lambda)}\right)\} \\ &s.t.\ \hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \Phi(\beta, \lambda) \ . \end{aligned} \tag{2}$$

Les méthodes classiques pour résoudre Problem (2) reposent souvent sur de l'optimisation à l'ordre zéro (c'est à dire qu'elles n'utilisent pas l'information du gradient) comme la *grid-search*, *random-search* [8], ou *l'optimisation bayésienne* [2, 9]. Cependant quand l'hyperparamètre est continu et que le chemin de régularisation $\lambda \mapsto \hat{\beta}^{(\lambda)}$ est (presque partout) différentiable, des méthodes d'optimisation du premier ordre peuvent être utilisées pour résoudre le problème d'optimisation à deux niveaux Problem (2). En effet en utilisant la dérivation de fonctions composées, le gradient de $\mathcal{L}$ en fonction de λ, aussi appelé *hypergradient*, s'écrit:

$$\nabla_\lambda \mathcal{L}(\lambda) = \hat{\mathcal{J}}_{(\lambda)}^\top \nabla \mathcal{C}(\hat{\beta}^{(\lambda)}) \ , \tag{3}$$

avec $\hat{\mathcal{J}}_{(\lambda)} \in \mathbb{R}^{p \times r}$ la Jacobienne de la fonction $\lambda \mapsto \hat{\beta}^{(\lambda)}$. La difficulté des méthodes du premier ordre pour résoudre Problem (2) est d'évaluer le gradient en Equation (3). Il y a trois algorithmes principaux de différenciation automatique pour calculer ce gradient: la différenciation implicite [4], la différenciation *backward* [6] et la différenciation *forward* [11]. Comme illustré dans la Figure 1, une fois que l'hypergradient à été calculé, Problem (2) peut être résolu en utilisant une descente de gradient par exemple. Avec un pas $\rho > 0$: $\lambda^{(t+1)} = \lambda^{(t)} - \rho \nabla_\lambda \mathcal{L}(\lambda^{(t)})$.

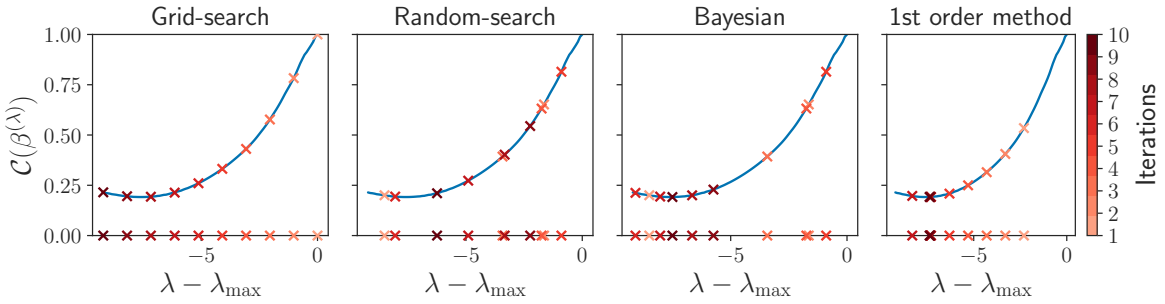


Figure 1: **Lasso CV sur le jeu de données *real-sim*.** Valeur de la fonction de validation croisée $\mathcal{C}(\beta^{(\lambda)})$ en fonction du paramètre de régularisation $\lambda \in \mathbb{R}$ pour plusieurs méthodes d'optimisation d'hyperparamètres pour le Lasso $\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + e^{\lambda}\|\beta\|_1$. Les croix représentent les 10 premières évaluations de chaque méthode.

2 Differentiation implicite

Le calcul de l'hypergradient en utilisant la différenciation implicite implique la résolution d'un système linéaire de taille $p \times p$ qui peut être difficile à résoudre lorsque p devient grand. Nous allons maintenant montrer que la différenciation implicite peut être utilisée pour calculer la Jacobienne de Problem (1). Les fonctions convexes non lisses en machine learning impliquent très souvent une structure particulière sur l'estimateur. Cette structure est portée par la notion de support généralisé que nous définissons ici:

Definition 1 (Support généralisé support) *Soit $\hat{\beta}$ une solution de Problem (1). Le* support généralisé $\hat{S} \subseteq [p]$ *est l'ensemble des indices $j \in [p]$ où g_j est différentiable en $\hat{\beta}_j$:*

$$\hat{S} \triangleq \{j \in [p] : \partial g_j(\hat{\beta}_j) \text{ est un singleton}\} \ . \tag{4}$$

De plus le type de problème Problem (1) peut être résolu en utilisant une descente de gradient proximale. Cette méthode de résolution itérative a deux caractéristiques intéressantes pour ce travail: 1. la descente de gradient proximale identifie le support après un nombre fini d'itération [5], 2. elle induit une équation de point fixe:

$$\hat{\beta}^{(\lambda)} = \operatorname{prox}_{\gamma g}\left(\hat{\beta}^{(\lambda)} - \gamma X^\top \nabla f(X\hat{\beta}^{(\lambda)})\right) \ . \tag{5}$$

Ces deux propriétés nous permettent de présenter un théorème pour faire de la différenciation implicite sur le problème générique Problem (1).

Theorem 2 (Differentiation implicite pour les problèmes non lisse) *On suppose que g_j est localement $\mathcal{C}^2$ pour tous les $j \in \hat{S}$ et que f est localement $\mathcal{C}^2$ autour de $X\hat{\beta}$. De plus on suppose que $-\nabla f(\hat{\beta}) \in \operatorname{ri}\left(\partial_1 g(\hat{\beta}, \lambda)\right)$ où* ri *est l'intérieur relatif du sous différentiel de g et que $X_{:\hat{S}}^\top \nabla^2 f(X\hat{\beta}) X_{:\hat{S}}$ est définie positive. Soit $\hat{\beta} \triangleq \hat{\beta}^{(\lambda)}$ une solution de Problem (1) et $\hat{S}$ son support généralisé. Alors la Jacobienne $\hat{\mathcal{J}}$ de Problem (1) est donné par la*

formule suivante: $\hat{z} = \hat{\beta} - \gamma \nabla X^\top f(X\hat{\beta})$, *et* $A \triangleq \mathrm{Id}_{\hat{S},\hat{S}} - \partial_1 \operatorname{prox}_{\gamma g}(\hat{z})_{\hat{S}} \left(\mathrm{Id}_{|\hat{S}|} - \gamma X_{:\hat{S}}^\top \nabla^2 f(X\hat{\beta}) X_{:\hat{S}} \right)$*:*

$$\hat{\mathcal{J}}_{\hat{S}^c} = \partial_2 \operatorname{prox}_{\gamma g}(\hat{z})_{\hat{S}^c} \quad ,$$
$$\hat{\mathcal{J}}_{\hat{S}} = A^{-1} \left(\partial_2 \operatorname{prox}_{\gamma g}(\hat{z})_{\hat{S}} - \gamma \partial_1 \operatorname{prox}_{\gamma g}(\hat{z})_{\hat{S}} X_{:\hat{S}}^\top \nabla^2 f(X\hat{\beta}) X_{:\hat{S}^c} \hat{\mathcal{J}}_{\hat{S}^c} \right) \quad .$$

Ce théorème prouve qu'il est possible d'utiliser la parcimonie induite par le support généralisé pour résoudre le système linéaire de manière efficace. En effet, le problème à résoudre est de taille $|\hat{S}| \times |\hat{S}|$ alors qu'il était de taille $p \times p$ pour les problèmes lisses. Dans beaucoup de cas pratiques, $|\hat{S}|$ est beaucoup plus petit que p, il suffit de penser au Lasso par exemple.

3 Optimisation d'hyperparamètres pour l'elastic net

Nous donnons maintenant un exemple de sélection d'hyperparamètre pour le modèle de l'elastic net utilisant une méthode de premier ordre et qui utilise Theorem 2 pour le calcul de l'hypergradient Equation (3). La sélection des valeur pour $\lambda \in \mathbb{R}^2$ est faite en utilisant une validation croisée à 5-blocs. Le problème à deux niveaux à résoudre est:

$$\begin{aligned} &\underset{\lambda=(\lambda_1,\lambda_2)\in\mathbb{R}^2}{\arg\min} \; \mathcal{L}(\lambda) = \tfrac{1}{n_{\text{fold}}} \sum_{i=1}^{n_{\text{fold}}} \|y^{\text{val}_i} - X^{\text{val}_i} \hat{\beta}^{(\lambda,i)}\|_2^2 \\ &s.t. \; \hat{\beta}^{(\lambda,i)} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{2n} \|y^{\text{train}_i} - X^{\text{train}_i}\beta\|_2^2 + e^{\lambda_1}\|\beta\|_1 + \frac{1}{2} e^{\lambda_2} \|\beta\|_2 \quad . \end{aligned}$$

La Figure 2 représente le critère de validation croisée en fonction du paramètre de régularisation (3 premières lignes) et en fonction du temps (dernière ligne). Les trois premières lignes montrent l'évaluation du critère pour la grid-search, une méthode bayésienne et notre méthode du premier ordre. Les croix de couleur les plus claires marquent les premières itérations. Sur tous les jeux de données, on peut remarquer que les méthodes du premier ordre sont plus rapides pour résoudre le problème d'optimisation et ainsi effectuer de la sélection d'hyperparamètres.

4 Conclusion

Dans ce travail, nous proposons une méthode de différentiation implicite pour la sélection des hyperparamètres dans le cadre de fonctions séparables et non lisses. Notre approche permet d'obtenir les hyperparamètres optimaux plus rapidement que les méthodes état de l'art surtout lorsque le nombre des hyperparamètres devient grand. La méthode peut s'appliquer à un grand nombre de modèles utilisés en apprentissage automatique comme le lasso, l'elastic net ou encore la SVM.

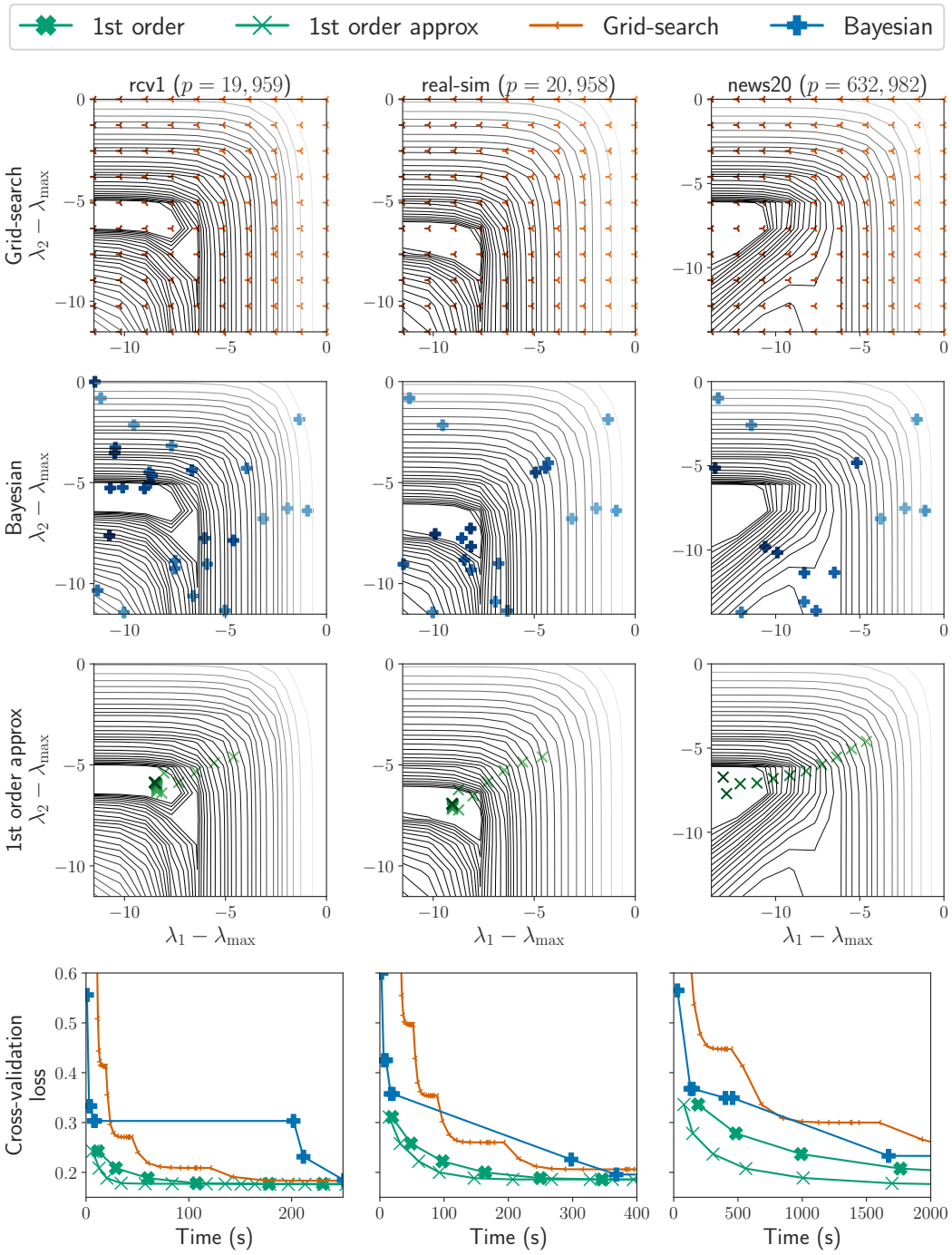


Figure 2: **Comparaison en temps pour l'elastic net.** Lignes de niveaux de la fonction de validation croisée 5-blocs en fontion des deux hyperparamètres λ_1 et λ_2 (sur les 3 premières lignes), et en fontion du temps (dernière ligne) pour les jeux de données *rcv1*, *real-sim* et *20news*.

References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[2] E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. 2010.

[3] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[4] J. Larsen, L. K. Hansen, C. Svarer, and M. Ohlsson. Design and regularization of neural networks: the optimal use of a validation set. In *Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, 1996.

[5] J. Liang, J. Fadili, and G. Peyré. Local linear convergence of forward–backward under partial smoothness. In *NeurIPS*, pages 1970–1978, 2014.

[6] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, pages 6–7, 1970.

[7] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. 1989.

[8] L. A. Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, 24:1337–1342, 1963.

[9] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *NeurIPS*, pages 2960–2968, 2012.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

[11] R. E. Wengert. A simple automatic derivative evaluation program. *Communications of the ACM*, 7(8):463–464, 1964.

[12] H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Sur les fonctions poids exponentiels et le phénomène de variation

Célestin C. Kokonendji[1], Aboubacar Y. Touré[2] & Rahma Abid[3]

[1] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, France.*
celestin.kokonendji@univ-fcomte.fr
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, France.*
aboubacar_yacouba.toure@univ-fcomte.fr
[3] *University Paris-Dauphine Tunis and Laboratory of Probability & Statistics of Sfax, Tunisia.*
rahma.abid@dauphine.tn

Résumé. Les lois exponentielles pondérées générales qui contiennent les lois exponentielles modifiées, sont largement utilisées dans les applications statistiques telles qu'en fiabilité. Nous étudions leurs fonctions poids exponentiels et des extensions à partir d'une loi de référence continue positive. Des propriétés et leurs relations avec le récent phénomène de variation sont établies. Des caractérisations, des opérations de pondération et de dualité sont établies. Enfin, des perspectives sont discutées.

Mots-clés. Dualité, fiabilité, loi pondérée, modèle exponentiel.

Abstract. General weighted exponential distributions including modified exponential ones are widely used with great ability in statistical applications, particularly in reliability. We investigate full exponential weight functions and their extensions from any nonnegative continuous reference distribution. Several properties and their connections with the recent variation phenomenon are then established. Characterizations, duality and weightening operations are set forward. Perspectives are discussed.

Keywords. Duality, Exponential model, Reliability, Weighted distribution.

1 Introduction

Le phénomène de variation a été introduit par Abid et al. (2020). Appelé indice de variation exponentielle ou indice de variation de Jørgensen pour la variable aléatoire (v.a.) continue positive Y sur $[0, +\infty[^k$, il est défini comme le rapport de la variance au carré de l'espérance. Plus précisément, cette quantité positive s'écrit

$$\mathrm{VI}(Y) := \mathrm{Var}Y/(\mathbb{E}Y)^2 \gtreqless 1. \tag{1.1}$$

Il peut être considéré comme le carré du coefficient de variation et est beaucoup utilisé en fiabilité. Voir, par exemple, Barlow et Proschan (1981) dans le sens du coefficient de variation. *L'indice de variation relative*, RVI, a également été introduit

comme le rapport de deux VIs en changeant la loi de référence exponentielle. Voir Kokonendji et al. (2020a) pour plus de détails sur le cas multivarié.

La fonction densité de probabilité (fdp) de la référence qui est la v.a. exponentielle $X \sim \mathcal{E}(\mu)$ de paramètre $\mu > 0$ est

$$f_X(x;\mu) = \mu \exp(-\mu x)\mathbb{1}_{[0,+\infty[}(x), \tag{1.2}$$

où $\mathbb{1}_A$ désigne la fonction indicatrice de A. Toujours équi-varié, sa moyenne et sa variance sont $1/\mu$ et $1/\mu^2$, respectivement. Il peut arriver que la variance de l'échantillon soit supérieure ou inférieure au carré de la moyenne de l'échantillon, que l'on appelle *sur-variation* et *sous-variation*, respectivement, par rapport à la loi exponentielle.

On peut définir une v.a. exponentielle pondérée de façon générale, désignée par X^w, à partir de la référence $X \sim \mathcal{E}(\mu)$ de (1.2) et une fonction poids mésurable positive $w : [0,+\infty[\to [0,+\infty[$ telle que

$$w(\cdot) f_X(\cdot;\mu) =: f_{X^w}(\cdot;\mu) \tag{1.3}$$

soit une fdp, appélée loi exponentielle pondérée (LEP). Plus pratiquement, si la fonction poids $w_0(\cdot)$ est telle que $1 \neq \mathbb{E}w_0(X) < \infty$, alors $w(\cdot)$ introduite en (1.3) devient

$$w(\cdot) = w_0(\cdot)/\mathbb{E}w_0(X), \tag{1.4}$$

qui est finalement auto-normalisée. Notons qu'à partir de (1.4), la fonction poids non normalisée $w_0(\cdot) \equiv w_0(\cdot;\boldsymbol{\alpha})$ peut dépendre d'un nombre strictement positif k de paramètres $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k) \in \Theta_k \subseteq \mathbb{R}^k$. Alors la fonction poids exponentiel $w(\cdot) \equiv w(\cdot;\mu,\boldsymbol{\alpha})$ de (1.3) dépend de $k+1$ paramètres. Les LEPs peuvent être considérées comme des lois exponentielles modifiées et fournissent une approche unifiée pour gérer à la fois la sur-variation et la sous-variation.

Ainsi, les objectifs fondamentaux de cette communication sont de fournir quelques propriétés générales pour les LEPs et leurs liaisons avec le phénomène de variation. Nous étudions les LEPs à deux paramètres dans le cadre des modèles exponentiels de dispersion (MEDs); par exemple, Jørgensen (1997). Le reste du document est organisé comme suit. Dans la section 2, nous énonçons plusieurs résultats liés à la représentation d'une loi continue sur la demi-droite positive comme une version pondérée d'une autre, en particulier de la loi exponentielle. Nous introduisons également la dualité ponctuelle entre deux LEPs. La section 3 présente une connexion entre les LEPs des MEDs et le phénomène de variation. Enfin, une conclusion avec une approche semi-paramétrique des estimations des LEPs sont exposées dans la section 4.

2 Résultats sur les lois exponentielles pondérées (LEPs)

Nous étalons plusieurs propriétés liées à la représentation d'une loi continue positive comme une loi pondérée par rapport à une loi de référence donnée. Les démonstrations

de tous les résultats de cette section et de la suivante se trouvent dans Kokonendji et al. (2020c) avec un certain nombre d'exemples récents de la littérature.

Théorème 2.1. *Soit Y une v.a. continue positive avec fdp $f_Y(\cdot)$, de support $\mathbb{S}_Y \subseteq [0, +\infty[$ et soit $X \sim \mathcal{E}(\mu)$ de (1.2) avec $\mu > 0$. Alors*

$$Y \overset{d}{=} X^w, \tag{2.1}$$

où $\overset{d}{=}$ représente l'égalité en loi et $w(\cdot)$ est la fonction poids exponentiel donnée par

$$w(x) = [\exp(\mu x)/\mu] f_Y(x), \quad \forall x \in \mathbb{S}_Y. \tag{2.2}$$

La proposition suivante montre le lien entre deux représentations d'une loi continue positive comme LEP. Il peut être utilisé pour démontrer l'unicité de (2.1).

Proposition 2.2. *Soit X_1 et X_2 deux v.a. exponentielles de paramètres μ_1 et μ_2 respectivement. Si une v.a. continue positive Y est telle que $Y \overset{d}{=} X_1^{w_1}$ avec $w_1(\cdot)$ définie en (2.2), alors $Y \overset{d}{=} X_2^{w_2}$ avec*

$$w_2(x) = (\mu_1/\mu_2) \exp[(\mu_2 - \mu_1)x] w_1(x) \mathbb{1}_{\mathbb{S}_Y}(x). \tag{2.3}$$

Le résultat suivant montre que toute v.a. continue positive à sa propre représentation exponentielle pondérée de (2.1), que nous appelons auto-décomposition en une LEP.

Corollaire 2.3. *Soit Y une v.a. continue positive avec $\mathbb{S}_Y \subseteq [0, +\infty[$ et telle que $\mathbb{E}Y < \infty$. Alors Y suit une LEP par rapport à la référence exponentielle de paramètre $1/\mathbb{E}Y$.*

Le Corollaire 2.3 représente une v.a. continue positive donnée comme une version pondérée de référence exponentielle de même moyenne. A titre d'exemples, nous illustrerons entre autres les cas des lois gamma, lognormale et Weibull à trois paramètres.

Voyons maintenant la commutativité de l'opération de pondération sous une référence plus générale que celle de la loi exponentielle.

Théorème 2.4. *Soit Y une v.a. continue positive et soit $w_1(\cdot)$ et $w_2(\cdot)$ deux fonctions poids positives telles que $0 < \mathbb{E}w_j(Y) < \infty$ avec $j = 1, 2$ et $0 < \mathbb{E}[w_1(Y)w_2(Y)] < \infty$. Alors*

$$(Y^{w_1})^{w_2} \overset{d}{=} Y^{w_1 w_2} \overset{d}{=} (Y^{w_2})^{w_1}.$$

Le concept de dualité introduit par Kokonendji et al. (2008, Section 3) pour les lois de Poisson pondérées (LPPs) est étendu ici à toute loi continue positive de référence donnée, y compris la loi exponentielle.

Définition 2.5. *Soit Y une v.a. continue positive sur $\mathbb{S}_Y \subseteq [0, +\infty[$, et soit $w_0(\cdot)$ et $w_0^*(\cdot)$ deux fonctions poids positives. Les deux versions pondérées correspondantes Y^{w_0} et $Y^{w_0^*}$ sont dites* `paires duales` *par rapport à Y si et seulement si*

$$w_0(y)\, w_0^*(y) = 1, \quad \forall y \in \mathbb{S}_Y.$$

Une conséquence pratique de la dualité pour les LEPs est que cette loi fournit la variation opposée (i.e., sur-variation pour sous-variation et inversement). Par exemple, la loi exponentielle $\mathcal{E}(\mu)$ est auto-duale car sa fonction poids est $w(x) = 1, \forall x \in [0, +\infty[$.

Enfin, si $Y \overset{d}{=} Z^w$ avec $w(\cdot)$ positive et $\mathbb{E}Y = \mathbb{E}Z$, alors à partir de (1.1) et des résultats précédents, l'indice de variation relatif $\mathrm{RVI}_Z(Y)$ de Y par rapport à Z satisfait :

$$\mathrm{RVI}_Z(Y) := \mathrm{VI}(Y)/\mathrm{VI}(Z) = \mathrm{Var}Y/\mathrm{Var}Z \gtreqless 1 \iff w(\cdot) \gtreqless 1. \tag{2.4}$$

3 LEPs des MEDs et phénomène de variation

Dans cette partie, nous établissons une connection (2.4) entre les VIs et les LEPs comme membres des MEDs positifs ; voir Jørgensen et Kokonendji (2011) et Kokonendji et al. (2020b).

La fdp d'un MED positif a la forme suivante :

$$f(x; \theta, \phi) = a(x; \phi) \exp[\theta x - K(\theta; \phi)] \mathbb{1}_{[0,+\infty[}(x), \tag{3.1}$$

avec $\phi > 0$ le paramètre de dispersion, $\theta \in \Theta \cap (-\infty, 0) \neq \emptyset$ le paramètre canonique, $a(x; \phi)$ la fonction de normalisation et $K(\theta; \phi)$ la fonction cumulante. Lorsque ϕ est fixé (p.ex., $\phi = 1$), on obtient les familles exponentielles naturelles (FEN) positive. Ainsi, la loi de référence exponentielle de (1.2) est une FEN avec fdp

$$f_X(x; \theta) = \exp[\theta x + \log(-\theta)] \mathbb{1}_{[0,+\infty[}(x), \tag{3.2}$$

$\mu = 1/(-\theta) = \mu(\theta)$ et $\theta < 0$ pour $X \sim \mathcal{E}(\theta)$.

Le résultat suivant découle du Théorème 2.1 dans le cas des MEDs positifs.

Proposition 3.1. *Tout MED positif avec fdp donnée en (3.1) est une LEP avec*

$$w(\cdot; \theta, \phi) = a(\cdot; \phi)/\mathbb{E}_\theta[a(X; \phi)] = a(\cdot; \phi) \exp[-K(\theta; \phi)]/(-\theta), \ \phi > 0, \theta < 0,$$

sa fonction poids exponentiel correspondante et $X \sim \mathcal{E}(\theta)$ de (3.2).

Nous énonçons maintenant une relation entre le phénomène de variation et les LEPs des MEDs.

Théorème 3.2. *Si X^w suit une LEP dans les MEDs de la Proposition 3.1 avec*

$$\Delta_\mu := \mu d^2/d\mu^2 \log \mathbb{E}_\mu[a(X; \phi)] - 2d/d\mu \log \mathbb{E}_\mu[a(X; \phi)] - \mu \left(d/d\mu \log \mathbb{E}_\mu[a(X; \phi)]\right)^2.$$

Alors

$$\Delta_\mu \gtreqless 0 \iff X^w \overset{\succ}{\underset{\prec}{\succeq}} X \sim \mathcal{E}(\mu);$$

i.e., Δ_μ est positive, nulle et négative si et seulement si X^w est sur-, équi- et sous-varié, respectivement, par rapport à $X \sim \mathcal{E}(\mu)$.

Pour établir ce résultat, nous aurons besoin du lemme technique suivant.

Lemme 3.3. *Si X^w suit une LEP dans les MEDs de la Proposition 3.1, alors pour $\mu = 1/(-\theta)$,*

$$\begin{aligned}\mathrm{Var}_\mu(X^w) &= (\mathbb{E}_\mu X^w)^2 + \mu^4 d^2/d\mu^2 \log \mathbb{E}_\mu[a(X;\phi)] - 2\mu^3 d/d\mu \log \mathbb{E}_\mu[a(X;\phi)] \\ &\quad -\mu^4 \left\{d/d\mu \log \mathbb{E}_\mu[a(X;\phi)]\right\}^2.\end{aligned}$$

Le phénomène de variation (1.1) pour toute LEP dans les MEDs sera plus pratique à travers la proposition suivante avec des illustrations directes de la Table 1, qui peut être envisagée dans le cadre des modèles linéaires généralisés. Cette proposition peut être vue comme un analogue du Corollaire 2.3.

Proposition 3.4. *Si X^w suit une LEP dans les MEDs de la Proposition 3.1 avec $m = \mathbb{E}X^w$ positive et $\phi V(m/\phi) = \mathrm{Var}_m(X^w)$, où $V(\cdot)$ est sa fonction variance unitaire, alors*

$$\mathrm{VI}_m(X^w) := \phi V(m/\phi)/m^2 \gtreqless 1 \iff X^w \succcurlyeq\preccurlyeq X \sim \mathcal{E}(1/m).$$

Table 1: Exemples des LEPs dans les MEDs en utilisant la Proposition 3.4.

Type(s) de LEPs	$V(m)$	$\phi V(m/\phi)/m^2$	Auteur(s)
Gamma ($p = 2$)	m^2	$1/\phi$	Morris (1982)
Gaussienne inverse ($p = 3$)	m^3	m/ϕ^2	Letac et Mora (1990)
Ressel-Kendall* ($q = 3$)	$m^2 + m^3$	$1/\phi + m/\phi^2$	Letac et Mora (1990)
Tweedie ($p > 1$)	m^p	$m^{p-2}\phi^{1-p}$	Jørgensen (1997)
Geometric Tweedie ($q > 1$)	$m^2 + m^q$	$1/\phi + m^{q-2}\phi^{1-q}$	Abid et al. (2020)

* Ressel-Kendall ≡ Gaussienne inverse géométrique.

4 Conclusion

Nous avons fait une représentation de toute loi continue positive comme une LEP avec des caractérisations et opérations de pondération.

À partir du résultat du Corollaire 2.3, nous pouvons poursuivre dans les approches non-paramétriques et semi-paramétriques de Marshall et Olkin (2007). En fait, pour estimer toute fdp continue positive $f : [0, +\infty[\to \mathbb{R}$ sous l'hypothèse non-paramétrique, nous considérons l'approche semi-paramétrique

$$f(\cdot) = w(\cdot) f_{\mathcal{E}}(\cdot;\theta) =: f_w(\cdot;\theta),$$

avec $f_w(\cdot;\theta)$ la part inconnue mais purement paramétrique et dépendant de θ, et $w(\cdot)$ la part non-paramétrique sur $[0, +\infty[$ pour θ fixé. Selon les lisseurs appropriés sur $[0, +\infty[$ (p.ex., Kokonendji et Somé, 2018), des travaux dans ce sens sont en cours. On peut voir Kokonendji et al. (2009, 2017) pour les lois de comptage. Enfin, on envisage aussi attaquer les cas multivariés des LEPs ainsi que des LPPs de Kokonendji et al. (2008).

Bibliographie

Abid, R., Kokonendji, C.C. and Masmoudi, A. (2020). Geometric Tweedie regression models for continuous and semicontinuous data with variation phenomenon, *AStA Advances in Statistical Analysis*, 104, 33-58.

Barlow, R.A. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing: Probability Models*, Silver Springs, Maryland.

Jørgensen, B. (1997). *The Theory of Dispersion Models*, Chapman and Hall, London.

Jørgensen, B. and Kokonendji, C.C. (2011). Dispersion models for geometric sums, *Brazilian Journal of Probability and Statistics*, 25, 263-293.

Kokonendji, C.C., Bonat, W.H. and Abid, R. (2020b). Tweedie regression models and its geometric sums for (semi-)continuous data, *WIREs Computational Statistics* ,12, e1496, `https://doi.org/10.1002/WICS.1496`

Kokonendji, C.C., Mizère, D. and Balakrishnan, N. (2008). Connections of the Poisson weight function to overdispersion and underdispersion, *Journal of Statistical Planning and Inference*, 138, 1287-1296.

Kokonendji, C.C., Senga Kiessé, T. and Balakrishnan, N. (2009). Semiparametric estimation for count data through weighted distributions, *Journal of Statistical Planning and Inference*, 139, 3625-3638.

Kokonendji, C.C. and Somé, S.M. (2018). On multivariate associated kernels to estimate general density functions, *Journal of the Korean Statistical Society*, 47, 112-126.

Kokonendji, C.C., Touré, A.Y. and Sawadogo, A. (2020a). Relative variation indexes for multivariate continuous distributions on $[0, \infty)^k$ and extensions, *AStA Advances in Statistical Analysis*, 104, 285-307.

Kokonendji, C.C., Touré, A.Y. and Abid, R. (2020c). On general exponential weight functions and variation phenomenon, *Sankhya A*, DOI: 10.1007/s13171-020-00226-z.

Kokonendji, C.C., Zougab, N. and Senga Kiessé, T. (2017). Poisson-weighted estimation by discrete kernel with application to radiation biodosimetry. In *Biomedical Big Data and Statistics for Low Dose Radiation Research* - Extended Abstracts Fall 2015, vol. **VII**, Part II, Chap.19, pp. 115-120 (Editors: Ainsbury, E.A., Calle, M.L., Cardis, E., Einbeck, J., Gómez, G., Puig, P.), Springer Birkhäuser, Basel.

Letac, G. and Mora, M. (1990). Natural real exponential families with cubic variance functions, *The Annals of Statistics*, 18, 1-37.

Marshall, A.W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*, Springer, New York.

Morris, C.N. (1982). Natural exponential families with quadratic variance functions, *The Annals of Statistics*, 10, 65-80.

Risk-sensitive learning for heterogeneous frameworks

Yassine Laguel [1]

[1] *Université Grenoble Alpes, Grenoble, France.*
yassine.laguel@univ-grenoble-alpes.fr

Résumé. En apprentissage distribué, l'hétérogenéité statistique est un défi fondamental soulevant des questions éthiques comme l'équité en terme de qualité de service offerte à une population d'utilisateurs. Dans ce travail, nous proposons Δ-FL, un cadre d'apprentissage collaboratif pour gérer des utilisateurs qui ne se conforment pas à la distribution de données moyenne de la population. En s'appuyant sur la théorie des mesures de risques, nous introduisons une fonction objective basée sur le superquantile pour concentrer l'apprentissage sur les utilisateurs les moins favorisés. Nous proposons un algorithme adapté à ce cadre distribué. Nous présentons des expériences numériques démontrant les bénéfices de notre modèle pour les utilisateurs. Cet exposé est basé sur un travail en collaboration avec Krishna Pillutla et Zaid Harchaoui(University of Washington) et Jérôme Malick(CNRS) [2].

Mots-clés. Superquantiles, Optimisation robuste, Apprentissage fédéré, Mesures de risques.

In distributed learning, statistical heterogeneity is a key issue raising fairness concerns on the quality of service over a population of users. In this work, we propose Δ-FL, a collaborative learning framework to handle heterogeneous client devices which do not conform to the population data distribution. Building upon the theory of risk measures, we introduce an objective function that focus on most disadvantaged users. We propose an algorithm adapted to this distributed setting. We present concrete numerical evidences of the benefits of our algorithm for disadvanted users. This talk is based on a joint work with Krishna Pillutla and Zaid Harchaoui (University of Washington) and Jérôme Malick (CNRS) [2].

Keywords. Superquantiles, Distributionally robust optimization, Federated Learning, Risk measures.

1 Introduction

The proliferation of mobile phones, wearables and edge devices has led to an unprecedented growth in the generation of user interaction data. Systems which tap into the power of this rich data while respecting the privacy of users are geared to lead the next generation of intelligent applications and devices. Such systems have naturally heterogeneous local data distribution: knowing *what* global model should be aimed for and and *how* to learn it

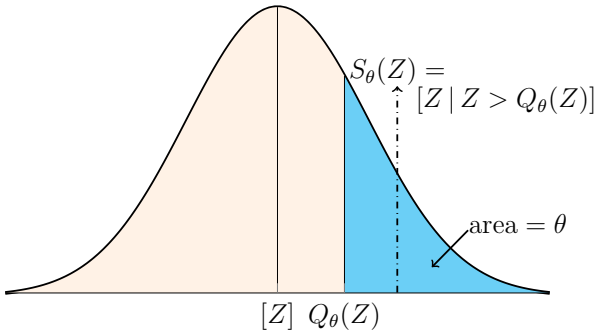


Figure 1: For a continuous random variable Z, drawing of $(1-\theta)$-quantile $Q_\theta(Z)$ and $(1-\theta)$-superquantile $S_\theta(Z)$, defined as an expectation.

become fundamental issues. In federated learning [1], client devices with privacy-sensitive data collaboratively learn a machine learning model under the orchestration of a central server, while keeping their data decentralized. This is achieved by pushing the computation to the devices while the server coordinates with the devices for aggregation of model updates. A key feature of federated learning is statistical heterogeneity, i.e., client data distributions are *not* identical. Each user has unique characteristics which are reflected in the data they generate. These characteristics are influenced by personal, cultural, and geographical factors. For instance, the varied use of language contributes to data heterogeneity in a next word prediction task.

Vanilla federated learning and its standard algorithm, FedAvg [6], aim to fit a model to the population distribution of the devices available for training. While this approach works for users who conform to the population, it is liable to fail on individuals who do not conform to the population, leading to poor user experience.

In this talk, we present risk measures [8] to train models that are more focused on disfavoured devices. More specifically, instead of minimizing an average of the local losses that measure the performance on each device, we propose to minimize the superquantile of such sequence. This has the effect of pushing optimization with respect to devices with a higher local loss. Optimization of such loss can be performed in large scale settings and has been well studied in the centralized setting [3]. We propose an algorithm to minimize such objective while satisfying the specific requirements of federated settings (keeping the data decentralized, allow only privacy preserving communications with the server, etc...).

2 A risk-sensitive model for federated learning

We consider a heterogeneous distributed setting with N training devices. We characterize each training device k by a probability distribution q_k over data and a weight $\alpha_k > 0$. For any model w belonging to some compact set $X \subset \mathbb{R}^d$, we measure the loss incurred by

w under data distribution q_k with a given function $f(w, q_k)$. We assume any test device, unseen during training, to have a distribution q_π that can be written as a mixture of the training distributions: there exists $\pi \in \Delta_{N-1}$ such that $q_\pi = \sum_{k=1}^{N} \pi_k q_k$. Our goal is to propose a global model $w^\star$ which (a) maintain good predictive power on test devices who conform to the population, and, (b) improve the predictive power on test devices who do not conform to the population.

Given the mixture distribution q_π of a test device, we define its *conformity* denoted $\text{conf}(q_\pi)$ as $\min_{k\in\{1,\dots,N\}} \alpha_k/\pi_k$. The conformity of a device is a summary of how close it is to the population. A test device with conformity $\theta \simeq 1$ closely conforms to the population distribution $q_\alpha = (\alpha_1, \dots, \alpha_n)$. In contrast, a test device with $\theta \simeq 0$ would be vastly different from the population distribution.

For a fixed parameter θ, we consider the robust optimization problem:

$$\min_{w\in X} \left[F_\theta(w) := \max_{\pi:\text{conf}(q_\pi)\geq\theta} f(w, q_\pi) \right]. \tag{1}$$

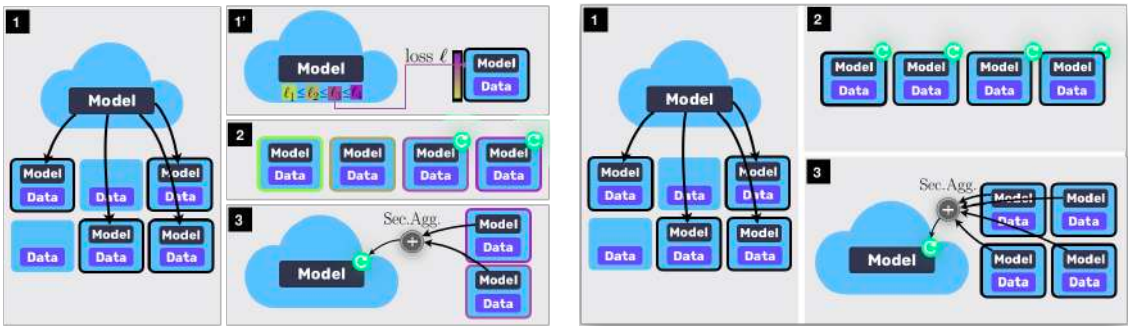
It is easy to show that this objective corresponds to the $(1-\theta)$-superquantile of the sequence of losses over the training devices. Note that when $\theta = 1$, we recover the standard average loss in federated learning minimized by FedAvg. In our work, we propose a second variant where

We use the dual form of this problem, which writes:

$$\min_{w\in X, \eta\in\mathbb{R}} \left[G_\theta(w, \eta) = \eta + \frac{1}{\theta} \sum_{k=1}^{N} \alpha_k [\max(f(w, q_k) - \eta, 0)] \right]. \tag{2}$$

A standard round of communication of our algorithm for solving (1) is consigned in Figure 2, next to FedAvg's one. Both consists of the following steps:

- **Step 1**: The server selects m client devices and broadcasts the global model to each selected device.
- **Step 1' (Δ-FL only)**: Each selected device computes the loss incurred by the global model on its local data and sends it to the server. Based on these losses, the server computes a threshold loss via the minimization of (2) with respect to η. It only keeps devices whose losses are larger than this threshold, and unselects the other devices.
- **Step 2**: Each selected device computes an update from its local data via local SGD.
- **Step 3**: The updates from selected devices are securely aggregated to update the server model.



(a) Δ-FL **(b)** Standard FedAvg

Figure 2: Description of one round of communication for our approach (Δ-FL) and the standard baseline in federated learning (FedAvg)

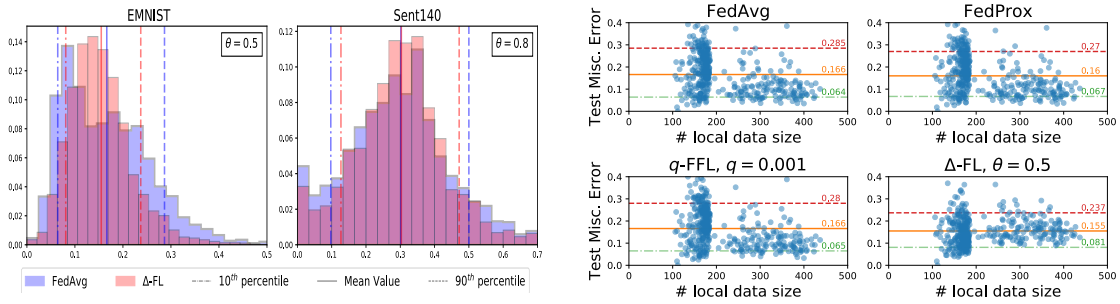
3 Theoretical and Experimental Results

3.1 Theoretical Analysis

We prove convergence of the algorithm in the smooth and strongly convex setting, with a convergence rate. Our analysis relies on classical assumptions, such as bounded variance of the local gradient estimators and bounded gradient dissimilarity among training devices. Further details can be found in our paper [2].

3.2 Numerical Experiments

We report experiments on two learning datasets: character recognition on EMNIST and sentiment analysis on tweets. We compare Δ-FL with the standard baseline FedAvg. We tried several non-convex models (ConvNet, LSTM, RNN). We track the misclassification error on each test device. To measure the performance on devices with low conformity, we reproduce the histogram of misclassification over test devices for both FedAvg and Δ-FL in the left part of Figure 3. We observe that our approach yields a thinner upper tail on this distribution and a lower variance, thus achieving its intended purpose. On the right part of Figure 3 we observe that improvement over the worst cases is achieved regardless of the local data size of the devices, while other established baselines such as FedProx [4], q-FFL [5] and AFL[7] all show lower performances on devices with smaller local datasets. More numerical evidences can be found in our paper [2].



(a) Histogram of misclassification error on test devices.

(b) Scatter plots of misclassification error on test devices against its data size for EMNIST.

Figure 3: Performance comparison between our approach Δ-FL and several established baselines on disfavoured devices

.

References

[1] Peter Kairouz et. al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14, 2021.

[2] Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. A superquantile approach for federated learning with heterogeneous devices. Accepted to the 55^{th} annual conference on Information Science and Systems (CISS 2021).

[3] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

[4] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems*, 2020.

[5] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *ICLR*, 2019.

[6] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, pages 1273–1282, 2017.

[7] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, 2019.

[8] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Estimation paramétrique de ruptures dans des données censurées à gauche

Clément Laroche [1] & Madalina Olteanu [2] & Fabrice Rossi [3]

[1] *Université Paris I- SAMM, Clement.Laroche@univ-paris1.fr*
[2] *Université Paris-Dauphine - Ceremade, Madalina.Olteanu@dauphine.psl.eu*
[3] *Université Paris-Dauphine - Ceremade, Fabrice.Rossi@dauphine.psl.eu*

Résumé. La phytopharmacovigilance a pour objectif de surveiller les effets indésirables des produits phytopharmaceutiques disponibles sur le marché et couvre notamment la contamination des milieux. Cette surveillance s'exerce notamment en recherchant des anomalies ou des ruptures dans les concentrations de substances actives de ces produits. La mesure d'une concentration chimique dépendant de la précision de la machine de mesure, les données observées sont soumises à une censure à gauche. Nous proposons ici une approche paramétrique permettant, en présence de données censurées et à seuil de censure connu, d'estimer le nombre et les positions des changements dans la concentration d'une substance. Les performances de cette approche sont comparées à celles d'une méthode de détection non-paramétrique, notamment sur des données réelles de concentrations du prosulfocarb en région Val de Loire.

Mots-clés. Détection de ruptures, censure à gauche, programmation dynamique, phytopharmacovigilance

Abstract. Pesticide effects are routinely monitored to detect potential health hazards. Pesticide residues in the environment are of particular interest. A natural monitoring strategy consists in looking for anomalies or change points in the concentration of active substances in a given environment. Due to the quantification limit of most chemical analyses, concentration observations are left censored which introduces some difficulties. We propose in this paper a method to detect change points, both in number and positions, in left censored concentration data that follow a parametric distribution (knowing the censoring threshold). The method is compared to a reference non parametric method on simulated and real world data (prosulfocarb concentration in Val de Loire).

Keywords. Change point detection, left censorship, PELT, pesticide monitoring.

1 Introduction

Le suivi de concentration de pesticides est un enjeu majeur des agences de sécurité environnementale et de santé publique. En France, il est rendu possible par la collecte de

mesures de concentration pour un grand nombre de substances, dans différents environnements, sur l'ensemble du territoire. Ces données, aujourd'hui publiques, nécessitent des méthodes d'analyse prenant en compte leurs spécificités [6]. Parmi celles-ci, on peut noter une grande hétérogénéité de collecte avec un rythme de relève irrégulier et une densité de collecte très variable en fonction de l'échelle géographique considérée.

De plus, les données de concentration présentent un phénomène de censure à gauche induit par les limites de précision des techniques d'analyse. Lorsque la concentration dans le prélèvement est trop faible, on observe la *limite de quantification* (LQ) de la technique employée et pas la véritable concentration sous-jacente.

Pour suivre les concentrations des substances actives, nous proposons de rechercher des ruptures dans ces valeurs, en adaptant les méthodes classiques [7] au cas de la censure à gauche. Il s'agit de fournir aux analystes des plages de concentration homogène et d'attirer l'attention sur des changements brutaux. Nous nous plaçons en outre dans un cadre paramétrique car les concentrations présentent souvent des distributions de type exponentiel.

La suite de l'article est organisée de la manière suivante. Nous présentons en section 2 le modèle de détection de ruptures ainsi que la procédure d'estimation des paramètres. La section 3 est consacrée à des expériences sur des données simulées. Nous concluons l'article par une section dédiée à une application sur des données réelles.

2 Détection de ruptures dans des données censurées à gauche

On suppose dans la suite que l'on dispose d'une série de réalisations $y_1, ..., y_n$ des variables aléatoires indépendantes $Y_1, ..., Y_n$. On notera par ailleurs $Y_{a:b} = (Y_a, ..., Y_b)$.

2.1 Modélisation

On suppose que les Y_i suivent des lois exponentielles censurées à gauche avec un seuil de censure commun, a. En pratique, ce seuil correspondant à la limite de quantification de l'analyse, est fixé et connu a priori. Les intensités des lois sont supposées constantes par morceaux. Plus précisément, on suppose qu'il existe K^* ruptures associées à $K^\star + 1$ intervalles définis par les instants $\boldsymbol{t}^\star = (0 = t_0^\star < t_1^\star < ... < t_{K^\star-1}^\star < t_{K^\star}^\star < t_{K^\star+1}^\star = n)$. Sur l'intervalle $]t_k^\star, t_{k+1}^\star]$, les Y_i concernées ont une intensité $\lambda_k^\star$.

On cherche à déterminer $K^\star$, $\boldsymbol{t}^\star$ et $\boldsymbol{\lambda}^\star = (\lambda_0^\star, ..., \lambda_{K^\star}^\star)$ à partir des observations. Pour ce faire, on utilise une approche classique de vraisemblance pénalisée, ce qui conduit au critère suivant

$$\tilde{\mathcal{C}}_{Y_{1:n}}(K, \boldsymbol{t}, \boldsymbol{\lambda}) = \sum_{k=0}^{K} W(Y_{(t_k+1):t_{k+1}}, \lambda_k) - \beta_n K \ , \tag{1}$$

où β_n désigne un terme de pénalité associé au rajout d'une nouvelle rupture, et $W(Y_{(t_k+1):t_{k+1}}) = \sum_{i=t_k+1}^{t_{k+1}} \ln f_{\lambda_k}(Y_i)$ la log-vraisemblance du segment $Y_{(t_k+1):t_{k+1}}$.

Le terme β_n peut être choisi de manière à obtenir les pénalités usuelles type AIC ou BIC, ou via des méthodes de calibration non-asymptotiques comme l'heuristique de pente [1].

L'estimateur de maximum de vraisemblance pénalisée est

$$(\hat{K}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{\lambda}}) = \arg \max_{K=1,\dots,K_{\max}, \boldsymbol{t}\in\mathcal{T}_K^{\Delta}, \boldsymbol{\lambda}\in\mathbb{R}_+^K} \tilde{\mathcal{C}}_{Y_{1:n}}(K, \boldsymbol{t}, \boldsymbol{\lambda}) \ , \tag{2}$$

où $\mathcal{T}_K^{\Delta} = \{\boldsymbol{t} = (0 = t_0 < t_1 < ... < t_{K-1} < t_K = n)\}$. Notons que K est contraint à être inférieur à une valeur $K_{\max}$ fixée par l'analyste sans que cela n'entraîne de perte de généralité. Par ailleurs, selon des arguments similaires à ceux dans [4], l'estimateur $(\hat{K}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{\lambda}})$ est asymptotiquement consistant.

2.2 Procédure d'estimation

Quand K et $\boldsymbol{t}$ sont fixés, l'additivité du critère permet de le maximiser par rapport à $\boldsymbol{\lambda}$ en travaillant segment par segment. On estime donc par maximum de vraisemblance l'intensité d'une loi exponentielle censurée à gauche. En raison de cette censure, il n'existe pas de formule explicite pour l'estimateur $\hat{\boldsymbol{\lambda}}$, on utilise donc une optimisation numérique (méthode de *Newton-Raphson* dans l'implémentation proposée ici). Une fois l'estimateur $\hat{\boldsymbol{\lambda}}$ obtenu, on effectue du *plug-in* pour calculer le score d'un segment.

Pour optimiser le critère par rapport à K et $\boldsymbol{t}$, on utilise l'algorithme *Pruned Exact Linear Time* (PELT [3]). Il permet d'obtenir une segmentation optimale efficacement en combinant programmation dynamique et élagage : le coût de calcul est en $\mathcal{O}(n)$. Notons que le critère retenu doit satisfaire certaines propriétés pour que PELT soit applicable, ce que nous avons vérifié dans le cas présent. L'une d'entre elles consiste en l'augmentation du score d'une séquence de données lors de l'introduction d'une rupture dans celle-ci.

3 Illustration sur données simulées

Dans un premier temps, la méthode paramétrique introduite ci-dessus sera comparée à l'état de l'art, et plus particulièrement à l'approche non-paramétrique *MultRank* décrite dans [5]. Cette dernière est inspirée par l'approche non-paramétrique usuelle basée sur la recherche de segments homogènes à partir d'un test sur les rangs, et on l'adapte au cas où de la censure est présente. On comparera en particulier la capacité des deux méthodes à détecter la présence d'une rupture dans les données. Lorsqu'on se place dans ce cadre, comparer l'approche non-paramétrique à l'approche paramétrique revient à utiliser un rapport de vraisemblance pour la dernière. Remarquons ici que réaliser un test de rapport de vraisemblance ou maximiser la vraisemblance pénalisée introduite dans l'Equation 2 pour $K_{max} = 1$ est équivalent, modulo le choix de la pénalité.

Les données simulées représentent $M = 2000$ échantillons de $n = 200$ réalisations d'une loi exponentielle. La moitié de ces échantillons ne contiendront pas de rupture, alors que l'autre moitié présentera une rupture en position $\frac{n}{2} = 100$. On notera λ_0 le paramètre de l'exponentielle du segment situé à gauche de la rupture et λ_1 celui du segment à droite. Les échantillons ne comportant pas de rupture seront générés selon une exponentielle de paramètre λ_0.

Les statistiques du test non-paramétrique et du test de rapport de vraisemblance seront calculées pour chaque échantillon, et permettront de calculer des courbes ROC et des aires sous la courbe associées, afin de comparer les performances des deux approches. A priori, les performances de l'approche paramétrique devraient être meilleures car le modèle est ici bien spécifié. Néanmoins, l'approche paramétrique devrait être plus sensible dans certains cas, et notamment en raison de la convergence très lente du rapport de vraisemblance et de la faible puissance du test [2] pour la distribution exponentielle.

Les résultats obtenus sur les données simulées sont présentés dans la Figure 1. Pour les deux méthodes, on calcule l'aire sous la courbe ROC, en faisant varier la nombre minimum d'observations avant une rupture. D'après ces résultats, si l'on contraint l'intervalle entre deux ruptures à contenir suffisamment d'observations (un dixième de la taille du signal ici pour $n = 200$), la méthode paramétrique obtient de meilleurs résultats. Par ailleurs, les performances des deux méthodes sont également comparées en fonction du seuil de censure. Dans le cas illustré ici, pour un seuil de censure égal à la médiane de la distribution dans le second segment, les résultats restent comparables, et les performances de la méthode paramétrique sont toujours meilleurs.

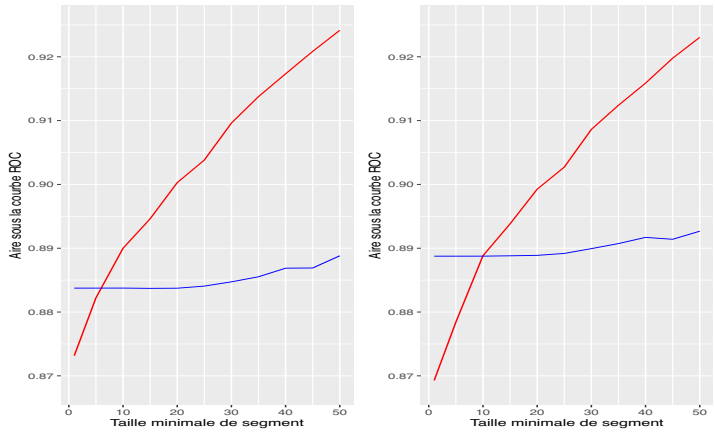


FIGURE 1 – Aire sous la courbe ROC. LR (likelihood ratio) : courbe rouge, MultRank : courbe bleue. $\lambda_0 = 4$, $\lambda_1 = 6$. Gauche : données sans censure. Droite : données avec censure $a = q_{50\%}$ d'une loi exponentielle de paramètre 6

4 Application sur données réelles

On étudie l'évolution de la concentration de prosulfocarbe entre les années 2007 et 2020. Le prosulfocarbe est un herbicide dont l'usage a été ré-autorisé en 2009. Depuis lors, les ventes de prosulfocarbe ont connu une explosion jusqu'à aujourd'hui, en passant de la dix-septième substance la plus vendue à la quatrième en 2017. Seulement les concentrations mesurées en région Centre-Val de Loire seront étudiées.

L'intérêt de ce cas d'étude réside dans le fait que la fenêtre d'observation temporelle dont nous disposons couvre des années où la substance était (normalement) absente des eaux, ainsi qu'une période de réapparition de cette substance active (son usage étant redevenu légal). Toutes les données utilisées dans cette section peuvent être téléchargées depuis l'adresse `http://www.naiades.eaufrance.fr/acces-donnees#/physicochimie`).

Les résultats obtenus via l'optimisation du critère pénalisé introduit dans l'Equation 2, ainsi que les résultats obtenus avec la méthode non-paramétrique *MultRank* sont illustrés dans la Figure 2. Pour la méthode paramétrique, on utilise une pénalité proportionnelle au BIC et une taille de minimale de segment égale à un dixième de la longueur signal total.

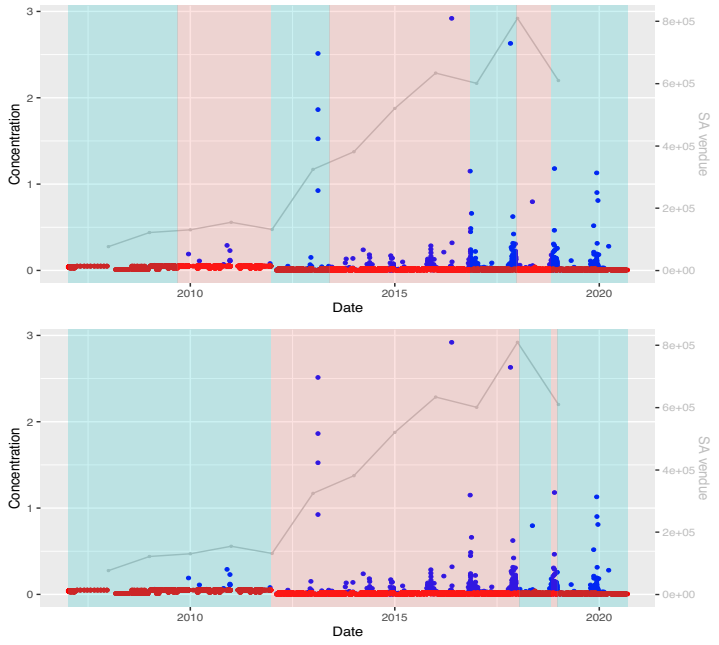


FIGURE 2 – Concentrations (en μg/L) de prosulfocarbe en Centre Val-de-Loire en fonction du temps. Les données censurées sont représentées en rouge. Haut : résultats de la méthode paramétrique. Bas : résultats de la méthode *MultRank*. La courbe grise représente le tonnage des ventes de prosulfocarbe par année en Centre Val de Loire.

Les ruptures détectées par la méthode paramétrique sont disposées de manière plus homogènes que celles détectées par *MultRank*. La taille minimale de segment ne permet pas de retrouver le résultat de *MultRank* (voir les deux ruptures formant le segment en 2019 du graphe du bas).

La méthode paramétrique positionne des ruptures dès que l'on observe une valeur élevée de concentration. Cela illustre la différence de robustesse entre les deux méthodes.

On peut remarquer que certaines positions de ruptures sont communes aux deux méthodes. Celle ayant eu lieu en 2018 coïncide avec un pic de concentration. Peu de temps précédant cette rupture se trouve la valeur la plus extrême de concentration du signal. Cette détection arrive lors de la plus grosse année de vente également.

Pour finir, la première rupture de *MultRank* en 2012 montre que les deux méthodes sont sensibles aux changements de LQ. Bien qu'elle corresponde au début de la croissance des ventes, on souhaiterait éviter de telles détections car elles ne correspondent pas à une augmentation dans les concentrations de prosulfocarbe (qui n'arrive qu'un an après). Cela s'explique plutôt par un changement de matériel de la part des laboratoires chargés de la mesure.

Références

[1] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2) :455–470, apr 2011.

[2] P. Haccou, E. Meelis, and S. van de Geer. The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic Processes and their Applications*, 27 :121–139, 1987.

[3] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500) :1590–1598, oct 2012.

[4] M. Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1) :79–102, sep 1999.

[5] A. Lung-Yut-Fong, C. Levy-Leduc, and O. Cappe. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 156(4) :133–162, 2015.

[6] J. Réty. Évaluation des risques liés aux résidus de pesticides dans l'eau de distribution : Contribution à l'exposition alimentaire totale. Technical report, Anses, 2013.

[7] Charles Truong. *Multiple change point detection – application to physiological signals.* Theses, Université Paris Saclay, November 2018.

Mélange de processus Gaussiens multi-tâches et prédictions cluster-spécifiques

Arthur Leroy [1] & Pierre Latouche [2] & Benjamin Guedj [3] & Servane Gey [4]

[1] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et arthur.leroy.pro@gmail.com*
[2] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et pierre.latouche@u-paris.fr*
[3] *Inria, France et University College London, United Kingdom et benjamin.guedj@inria.fr*
[4] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et servane.gey@u-paris.fr*

Résumé. La communication proposée porte sur l'analyse de données fonctionnelles et la définition de modèles de processus Gaussiens (GP) multi-tâches pour traiter simultanément les problèmes de régression et de classification non-supervisée. L'algorithme Magma, issu d'un travail antérieur, permet la modélisation de multiples séries temporelles asynchrones supposées partager de l'information commune, offrant une amélioration drastique des performances comparées à la régression GP traditionnelle, ainsi qu'une pleine prise en compte de l'incertitude. Nous introduisons une extension de ce modèle permettant la définition d'un mélange de GPs multi-tâches ajoutant un aspect clustering à cette approche. L'apprentissage des hyper-paramètres d'un tel modèle repose sur la définition de distributions variationnelles, la vraisemblance n'étant pas calculable directement, permettant de conserver une formulation explicite des lois a posteriori. Des formules analytiques sont également dérivées pour la prédiction de temps non-observés. L'algorithme MagmaClust ainsi obtenu permet à la fois d'identifier des structures de groupes dans un ensemble de courbes et offre des prédictions cluster-spécifique encore améliorées par rapport à Magma. Cette approche a été implémentée et expérimentée sur différents jeux de données simulés, offrant des performances remarquables tant sur les aspects de clustering que de prédiction. Une application sur données réelles est également proposée via l'étude et la prédiction future des courbes de performances de jeunes nageurs français.

Mots-clés. Processus Gaussiens, apprentissage multi-tâche, clustering de courbes, méthodes variationnelles

Abstract. The present work is dedicated to the analysis of functional data and the definition of multi-task Gaussian processes (GP) models for simultaneously dealing with regression and clustering. The algorithm Magma, from a previous work, enables modelling multiple asynchronous time series, assumed to share information, offering a remarkable improvement in performances compared to standard GP regression, along

with a thorough quantification of uncertainty. An extension of this work is proposed from the definition of a multi-task GPs mixture, which enriches the previous approach with a clustering aspect. Learning the hyper-parameters in such model lies on the definition of variational distributions, since likelihood is not available directly, allowing us to maintain explicit posterior distributions. In addition, analytical formulas are derived for prediction of unobserved timestamps. The resulting algorithm, MagmaClust, offers a group-structure identification within a set of curves as well as enhanced predictions compared to Magma. This approach has been implemented and tested on several simulated datasets, exhibiting noticeable performances both on clustering and prediction tasks. A real data application, focusing on the study and forecast of future performance curves for young french swimmers, is proposed as well.

Keywords. Gaussian Processes, multi-task learning, curve clustering, variational inference

1 Contexte

Le cadre des processus Gaussiens (Rasmussen and Williams, 2006) offre une modélisation élégante pour traiter le cas des données fonctionnelles, mais souffre toutefois de limitations lorsque les points d'observations sont peu nombreux et/ou mal répartis sur le domaine d'étude. Cependant, la définition de modèles multi-tâches (Caruana, 1997), autorisant le partage d'informations, permet de tirer le meilleur parti de situations où de multiples séries temporelles, présentant des caractéristiques communes, sont observées. Une approche classique pour définir un modèle de GPs multi-tâche a été introduit dans Bonilla et al. (2008) en définissant une structure de covariance particulière, composée de deux matrices, représentant respectivement les covariances entre les individus et entre les tâches. Toutefois, tant sur le point de la complexité algorithmique, de l'impossibilité de gérer des observations asynchrones, que sur des capacités prédictives raisonnablement limitées, cette méthode reste non optimale dans de nombreuses applications. Plus récemment, un algorithme du nom de Magma a été proposé (Leroy et al., 2020b) pour traiter l'entrainement et la prédiction d'une nouvelle formulation de modèles de GPs multi-tâches. L'originalité de cette approche repose sur l'introduction d'un processus moyen, commun à tous les individus, qui, une fois estimé, embarque une information partagée fournissant une moyenne a priori pré-entrainée avant même la prédiction. Les performances prédictives se trouvent être grandement améliorées, notamment loin des points d'observations, tout en conservant une complète quantification de l'incertitude et une gestion naturelle des données observées irrégulièrement d'une courbe à l'autre.

2 Modèle et inférence

Le travail ici présenté (Leroy et al., 2020a) s'inscrit dans la continuité de cette approche, en proposant une généralisation du modèle précédent à l'aide d'un mélange de GPs, permettant d'identifier une éventuelle structure de groupe dans les multiples tâches d'entrainement. En effet, pour certains jeux de données, l'hypothèse d'un unique processus central sous-jacent peut être trop restrictive. Ainsi, pour une donnée fonctionnelle y_i associée au i-ème individu appartenant au k-ième groupe, le modèle génératif se définit comme suit :

$$y_i = \mu_k + f_i + \epsilon_i,$$

où μ_k est un GP spécifique au k-ème groupe, alors que f_i et ϵ_i représentent un GP et un bruit gaussien, tous deux spécifiques à l'individu i. Une formulation hiérarchique équivalente, comme donnée ci-dessous pour tout vecteur de temps d'observation $\mathbf{t}$, permet de mieux comprendre en quoi les processus μ_k définissent les moyennes de chacun des clusters:

$$p\left(y_i(\mathbf{t}) \mid \mu_k(\mathbf{t})\right) = \mathcal{N}\left(y_i(\mathbf{t});\ \mu_k(\mathbf{t}), \mathbf{\Psi}_{\theta_i,\sigma_i^2}(\mathbf{t},\mathbf{t})\right),\ \forall i, \forall \mathbf{t},$$

où $\mathbf{\Psi}_{\theta_i,\sigma_i^2}$ désigne la structure de covariance associée à l'individu i. Ce nouveau modèle dépend également d'une variable multinomiale latente Z_i, contrôlant l'appartenance des individus à chaque cluster. Dans cette approche, il est à présent nécessaire d'estimer les hyper-paramètres des noyaux de covariance, conjointement des lois hyper-posterior des processus μ_k et des variables Z_i. Les dépendances a posteriori entre ces dernières quantités poussent à introduire un algorithme Variationnel EM (VEM) (Attias, 2000) pour l'entrainement, où les hyper-paramètres sont obtenus par maximisation de l'ELBO via l'algorithme d'optimisation *L-BFGS-B* (Morales and Nocedal, 2011). Nous dérivons des lois variationnelles approximées, dont les expressions analytique permettent leur utilisation ultérieure dans de formules de prédiction GP. Un algorithme EM est également établi pour estimer les hyper-paramètres associés à un nouvel individu, partiellement observé, ainsi que ses probabilités d'appartenance aux différents clusters. Par intégrations successives sur les processus moyens μ_k, puis sur les Z_i, une loi a posteriori de mélange gaussien multi-tâche peut être déduite, définie comme une somme pondérée de prédictions GP cluster-spécifiques.

3 Expériences et résultats

Nous illustrons au travers de simulations les avantages d'une telle approche et son intérêt lorsque les données présentent des structures de groupes. Par exemple, la Figure 2 propose une comparaison sur un même jeu de données entre la régression GP classique, l'algorithme MAGMA, et notre nouvel approche MAGMACLUST, pour prédire des points non observés (rouge) à partir d'observations (noir), aidé par les données issues des individus d'entrainement (colorés en arrière plan). Les performances sur les aspects de clus-

tering sont évaluées sur la Figure 1, et celles-ci dépassent nettement celles d'alternatives usuelles de la littérature. L'algorithme a également été appliqué dans le cadre d'une étude des courbes de progressions de jeunes nageurs, issues de données de la fédération française de natation. Ce travail a permis d'indentifier différents profils de progression parmi les individus ainsi qu'une prédiction probabiliste fiable des performances futures pour chaque sportif.

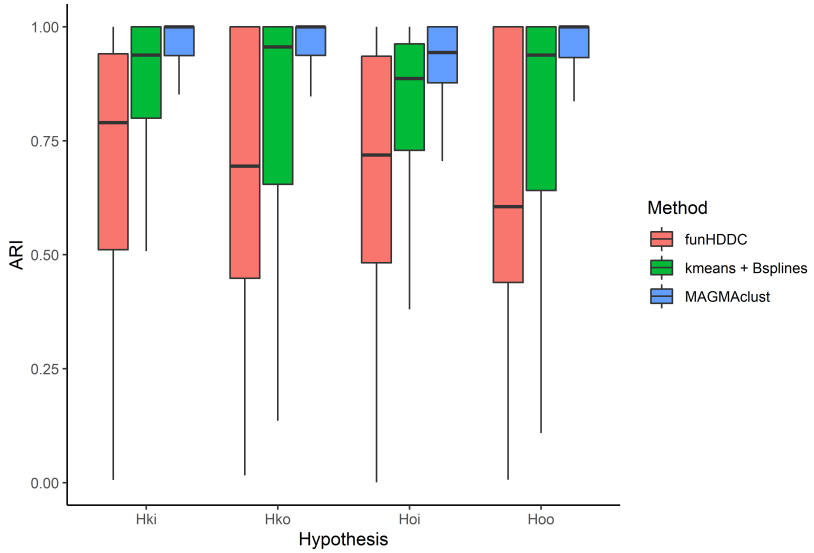


Figure 1: Valeurs des Adjusted Rand Index (ARI) entre les vrais clusters et les partitions estimées par les algorithmes kmeans, funHDDC, et MagmaClust. Le vrai nombre de groupes est spécifié dans chaque méthode et le ARI est calculé sur 100 jeux de données simulés selon 4 hypothèses différentes.

Bibliographie

Bibliographie

H. Attias. A Variational Baysian Framework for Graphical Models. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian Process Prediction. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 153–160. Curran Associates, Inc., 2008.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. Cluster-Specific Predictions with Multi-Task Gaussian Processes. *PREPRINT arXiv:2011.07866 [cs, LG]*, Nov. 2020a.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. MAGMA: Inference and Prediction with Multi-Task Gaussian Processes. *PREPRINT arXiv:2007.10731 [cs, stat]*, July 2020b.

J. L. Morales and J. Nocedal. Remark on algorithm L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 38(1):7:1–7:4, Dec. 2011. ISSN 0098-3500. doi: 10.1145/2049662.2049669.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9.

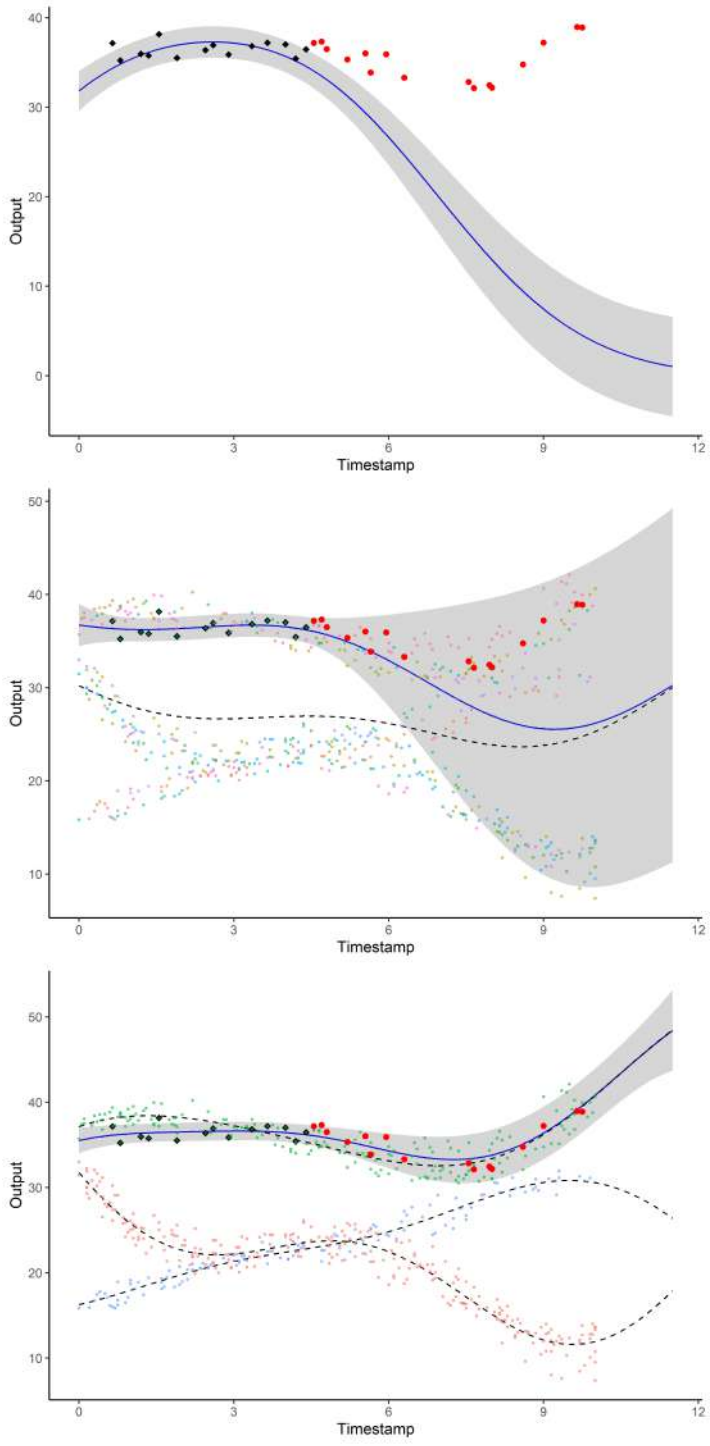


Figure 2: Courbes prédictives (bleu) et intervalles de crédibilité à 95% associés (gris) pour la régression GP (haut), MAGMA (milieu) et MAGMACLUST (bas). Les lignes pointillées représentent le pamarètre de moyenne de chaque processus moyens μ_k. Les points observés sont en noir, les points de test à prédire sont en rouge. Les points colorés en arrière plan sont issus des individus de la base d'entrainement.

A text based deep latent variable approach for missing rating imputation

Dingge Liang [1] & Marco Corneli [1,2] & Charles Bouveyron [1] & Pierre Latouche [3]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné*
[2] *Center of Modelling, Simulation and Interactions (MSI)*
[3] *Université de Paris, Laboratoire MAP5*

Abstract. We introduce a deep latent recommender system (deepLTRS) in order to provide users with high quality recommendations based on observed user ratings *and* texts of product reviews. The underlying motivation is that, when a user scores only a few products, the texts used in the reviews represent a significant source of information. Using this information can alleviate data sparsity, thereby enhancing the predictive ability of the model. Our approach adopts a variational auto-encoder architecture as a generative deep latent variable model for both the ordinal matrix, encoding users scores about products, and the document-term matrix, encoding the reviews. Moreover, different from unique user-based or item-based models, deepLTRS assumes latent representations for both users and products. An alternated user/product mini batch optimization structure is proposed to jointly capture user and product preferences. Numerical experiments on simulated and real-world data sets demonstrate that deepLTRS outperforms the state-of-the-art, in particular in context of extreme data sparsity.

Keywords. Recommender Systems, Learning Generative Models, Matrix Completion

1 Introduction and related work

In the current era of information explosion, recommendation systems have become central tools in a wide range of applications ranging from e-commerce to the global positioning of IoT devices. Examples of recommended objects include movies, songs, books, as well as restaurants to name just a few. At the core of the research in recommendation systems, we point out the collaborative filtering [Herlocker et al., 2000], content-based filtering [Pazzani and Billsus, 2007] and hybrid methods [Burke, 2002] which have been widely used to complete a matrix of user ratings about products based on the observed entries. In this paper, we consider the problem of completing a large and extremely sparse user/product matrix, when text reviews are available.

A long series of techniques have been proposed in the literature to address the matrix completion problem. On the one hand, most algorithms have been proposed on the basis of the sole knowledge of ratings. The HPF [Gopalan et al., 2015] model assumes that the observed rating matrix is drawn from a Poisson distribution with latent user preferences

and latent item attributes as parameters. It combines a sparsity model with a single response model. More recently, CCPF [Basbug and Engelhardt, 2017] was introduced by coupling a hierarchical Poisson factorization with an arbitrary data-generating model among three different methods: mixture models, linear regression and matrix factorization. Another set of recommender systems exploit both ratings and texts to improve predictions. The HFT [McAuley and Leskovec, 2013] combines latent rating factors with latent review topics by maximizing a penalized log-likelihood where the first term accounts for rating distribution and the second penalty term accounts for the words distribution over latent topics. However, HFT suffers from the limitation that the number of latent factors should be equal to the number of latent topics. The ALFM [Cheng et al., 2018] breaks this limitation by associating latent factors with different aspects. Each aspect is represented as a probability distribution of latent topics. The overall rating is computed through a linear combination of all the aspect ratings to achieve good performance.

Apart from models mentioned above, several deep-learning based methods have been proposed recently. For instance, DeepCoNN [Zheng et al., 2017] uses CNNs to learn representations of users and products from reviews and a regression layer is subsequently introduced for the prediction of ratings. However, DeepCoNN assumes that reviews are available only in the training phase. As an extension of DeepCoNN, the TransNet model was introduced in [Catherine and Cohen, 2017] with an additional layer that allows the model to also generate approximate comments during test and helps the model improve prediction performance.

In order to both improve the robustness to data sparsity and the interpretability of recommendations, we introduce here the deepLTRS for the rating matrix completion. It aims at accounting for both observed ratings and the textual information collected in product reviews. DeepLTRS extends the probabilistic matrix factorization [Mnih and Salakhutdinov, 2008] by relying on recent auto-encoding extensions [Srivastava and Sutton, 2017; Dieng et al., 2019] of latent Dirichlet allocation [Blei et al., 2003, LDA].

2 Generative model of deepLTRS

In this work, we consider data sets involving M users who are scoring and reviewing P products. Such data sets can be encoded by two matrices: **i)** an ordinal data matrix $Y \in \mathbb{N}^{M \times P}$, with Y_{ij} accounting for the *score* that the i-th user assigns to the j-th product. When a score is assigned, it takes values in $\{1, \ldots, H\}$ with $H > 1$; **ii)** a document-term matrix (DTM) W encoding the *reviews* that users write about products. By storing all the different vocables employed by the users into a *dictionary* of size V, the v-th entry of $W^{(i,j)}$, denoted by $W_v^{(i,j)}$, is the number of times that the word v appears into the review of the j-th product given by the i-th user. The document-term matrix W is obtained by concatenation of all the row vectors $W^{(i,j)}$.

It is now assumed that both users and products have latent representations in a low-

dimensional space $\mathbb{R}^D$, with $D \ll \min\{M, P\}$.

Ratings. The following generative model is now considered for the ratings:

$$Y_{ij} = \langle R_i, C_j \rangle + b_i^u + b_j^p + \epsilon_{ij}, \forall i = 1, ..., M, \forall j = 1, ..., P, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the standard scalar product and b_i^u, b_j^p are two unknown real parameters accounting for biases specific to users and products respectively. Finally, the residuals ϵ_{ij} are assumed to be i.i.d. normally distributed random variables, with zero mean and unknown variance η^2 as $\epsilon_{ij} \sim \mathcal{N}(0, \eta^2)$.

In the following, R_i and C_j are seen as random vectors, such that $R_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_D)$, $\forall i$ and $C_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_D)$, $\forall j$, with R_i and C_j assumed independent. The unbiased version of this model (i.e. with $b_i^u = b_j^p = 0$) is the well known probabilistic matrix factorization.

Reviews. We now extend the generative model to account for the document-term matrix W. Following the LDA model, each document is drawn from a mixture distribution over a set of K latent topics. The topic proportions in the document are denoted by a vector lying in the $K-1$ simplex. In deepLTRS, we assume that the topic proportions $\theta_{ij} \in [0,1]^K$, with $\sum_{k=1}^K \theta_{ij} = 1$. Moreover, they follow

$$\theta_{ij} = \sigma(f_\gamma(R_i, C_j)), \tag{2}$$

where $f_\gamma : \mathbb{R}^{2D} \to \mathbb{R}^K$ is a continuous function approximated by a neural network parametrized by γ and $\sigma : \mathbb{R}^K \to \mathbb{R}^K$ denotes the softmax function.

As in LDA, each document $W^{(i,j)}$ is seen as a vector in $\mathbb{N}^V$ (we recall that V is the dictionary size) obtained as

$$W^{(i,j)} | \theta_{ij} \sim \text{Multinomial}(L_{ij}, \beta\theta_{ij}), \tag{3}$$

where L_{ij} is the number of words in the review $W^{(i,j)}$ and $\beta \in [0,1]^{V \times K}$ is the matrix whose entry β_{vk} is the probability that vocable v occurs in topic k. By construction, $\sum_{v=1}^V \beta_{vk} = 1, \forall k$. In addition, conditionally to the vectors θ_{ij}, all the reviews $\{W^{(i,j)}\}_{i,j}$ are independent random vectors.

Finally, we emphasize that Y_{ij} and $W^{(i,j)}$ are *not* assumed to be independent. Instead, we described the above framework in which the dependence between them is completely captured by the latent embedding vectors R_i and C_j.

3 Variational auto-encoding inference

This section now details the auto-encoding variational inference procedure. Let us denote by $\Theta = \{\eta^2, \gamma, \beta, b^u, b^p\}$ the set of the model parameters introduced so far. A natural inference procedure would consist in looking for $\hat{\Theta}_{ML}$ maximizing the (integrated)

log-likelihood of the observed data (Y, W). Unfortunately, this quantity is not directly tractable and we rely on a variational lower bound to approximate it.

Let us consider a joint distribution $q(\cdot)$ over the pair (R, C) of all $(R_i)_i$ and $(C_j)_j$. Thanks to the Jensen inequality, it holds that

$$\begin{aligned}\log p(Y, W|\Theta) \geq & \mathbb{E}_{q(R,C)}\left[\log \frac{p(Y, W, R, C|\Theta)}{q(R, C)}\right] = \mathbb{E}_{q(R,C)}\left[\log p(W, Y|R, C, \Theta) + \log \frac{p(R, C)}{q(R, C)}\right] \\ = & \mathbb{E}_{q(R,C)}\left[\log p(W|R, C, \beta)\right] + \mathbb{E}_{q(R,C)}\left[\log p(Y|R, C, \gamma, \eta^2, b_u, b_p)\right] - D_{KL}(q(R, C)||p(R, C)),\end{aligned} \tag{4}$$

where D_{KL} denotes the Kullback-Leibler divergence between the variational posterior distribution of the latent row vectors $(R_i)_i, (C_j)_j$ and their prior distribution. The above inequality holds for every joint distribution $q(\cdot)$ over the pair (R, C). In order to deal with a tractable family of distributions, the following *mean-field* assumption is made

$$q(R, C) = q(R)q(C) = \prod_{i=1}^{M}\prod_{j=1}^{P} q(R_i)q(C_j). \tag{5}$$

Moreover, since R_i and C_j follow Gaussian prior distributions, $q(\cdot)$ is assumed to be as follows

$$q(R_i) = g(R_i; \mu_i^R := h_{1,\phi}(Y_i, W^{(i,\cdot)}), S_i^R := h_{2,\phi}(Y_i, W^{(i,\cdot)})), \tag{6}$$

and

$$q(C_j) = g(C_j; \mu_j^C := l_{1,\iota}(Y^j, W^{(\cdot,j)}), S_j^C := l_{2,\iota}(Y^j, W^{(\cdot,j)})), \tag{7}$$

where $g(\cdot; \mu, S)$ is the pdf of a Gaussian multivariate distribution with mean μ and variance S. The two matrices S_i^R and S_j^C are assumed to be diagonal matrices with D elements. Moreover, Y_i (respectively Y^j) denotes the i-th row (column) of Y, $W^{(i,\cdot)} := \sum_j W^{(i,j)}$ corresponds to a document concatenating all the reviews written by user i and $W^{(\cdot,j)} := \sum_i W^{(i,j)}$ corresponds to all the reviews about the j-th product. The functions $h_{1,\phi}$ and $h_{2,\phi}$ encode elements of $\mathbb{R}^{P+V}$ to elements of $\mathbb{R}^D$. Similarly, $l_{1,\iota}$ and $l_{2,\iota}$ encode elements of $\mathbb{R}^{M+V}$ to elements of $\mathbb{R}^D$. These functions are known as the network *encoders* parametrized by ϕ and ι, respectively.

Thanks to Eqs. (1)-(3)-(5)-(6)-(7) and by computing the KL divergence in Eq. (4), the *evidence lower bound* (ELBO) on the right hand side of Eq. (4) can be further developed as follows:

$$\begin{aligned}\text{ELBO}(\overline{\Theta}) = & \sum_{i,j}\left(\mathbb{E}_{q(R_i,C_j)}\left[-\frac{1}{2}\left(\frac{(Y_{ij} - (R_i^T C_j + b_u^i + b_p^j))^2}{\eta^2} + \log \eta^2\right)\right]\right) \\ & + \sum_{i,j}\left(\mathbb{E}_{q(R_i,C_j)}\left[\left(W^{(i,j)}\right)^T \log\left(\beta\sigma(f_\gamma(R_i, C_j))\right)\right]\right) - \sum_i\left[-\frac{1}{2}\left(tr(S_i^R) + (\mu_i^R)^T\mu_i^R - D - \log|S_i^R|\right)\right] \\ & - \sum_j\left[-\frac{1}{2}\left(tr(S_j^C) + (\mu_j^C)^T\mu_j^C - D - \log|S_j^C|\right)\right] + \xi\end{aligned} \tag{8}$$

where now $\overline{\Theta} := \{\eta^2, \gamma, \beta, \phi, \iota, b_u, b_p\}$ denotes the set of the model *and* variational parameters and ξ is a constant term that includes all the elements not depending on $\overline{\Theta}$.

The deep view of deepLTRS is shown in Figure 1. The *encoders* $h_{1,\phi}, h_{2,\phi}$ and $l_{1,\iota}, l_{2,\iota}$ map the observed data from $\mathbb{R}^{P+V}$ and $\mathbb{R}^{M+V}$, respectively, to the variational parameters in a lower dimension $\mathbb{R}^D$. Symmetrically, the role of *decoder* is played by: **i)** RC^T, the matrix product of R and C, that maps the lower dimension representations to the "reconstructed" ordinal data matrix $\hat{Y}$; **ii)** β, which maps the topic proportions from $\mathbb{R}^K$ into vectors in $\mathbb{R}^V$ (the "reconstructed" rows of $\hat{W}$).

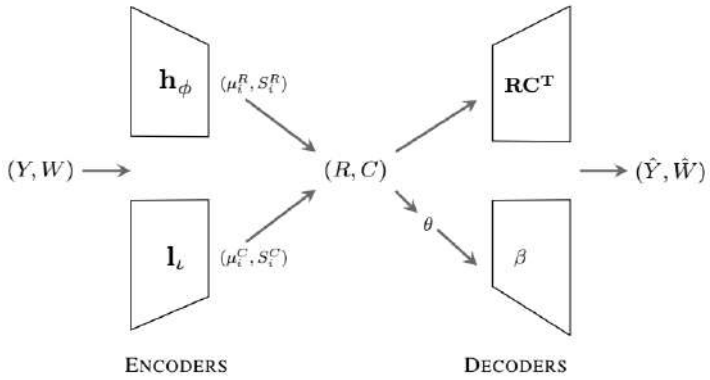


Figure 1: A deep-learning-like model view of DeepLTRS.

4 Application on real-world data

We now consider applying deepLTRS to real-world data sets consisting of different product reviews from Amazon. The data sets can be downloaded freely on the dedicated websites[1][2]. Five independent runs of the algorithm were performed. For each run, we randomly selected 80%, 10% and 10% of the data as the training, validation and test set. We trained our model for 100 epochs using Adam optimizer, with a learning rate of $2e^{-3}$.

Table 1 presents the test RMSE for deepLTRS and its competitors on the predicted ratings for Amazon data sets. Reported test RMSE is obtained when the RMSE on the validation set was the lowest, as for all methods. First of all, both HPF and CCPF models only considered the user rating information. By replacing the single Poison distribution in HPF with a mixture model, CCPF has made great improvements in RMSE. Next, the remaining four methods all consider ratings and reviews. Among them, TransNet and deepLTRS are deep-learning based models. It can be seen that, in general, ALFM and deepLTRS always have better performance than HFT and TransNet.

[1]Amazon Fine Food reviews `https://snap.stanford.edu/data/web-FineFoods.html`
[2]Amazon Product data `https://jmcauley.ucsd.edu/data/amazon/`

It is worth mentioning that deepLTRS outperforms ALFM on two data sets, Fine food and Patio, while ALFM has better performance on the other two data sets since ALFM introduces the average of all ratings to the formula in the score generation phase. When most of the scores of the experimental data are very positive, for example, a lot of scores are equal to 4, ALFM can achieve good results thanks to this average bias parameter. However, if the score distribution of the data is more scattered, ALFM can not perform well. To confirm this fact, we built a data set with ratings simulated in an interval [1,4] and texts extracted from four BBC news. As shown in Figure 2, ALFM makes all predictions concentrated near the average, which is not consistent with the simulation setup.

Table 1: Test RMSE on Amazon data sets.

Data sets	HFT	HPF	CCPF-PMF
Fine Food	1.4477 (±0.0465)	2.9528 (±0.0144)	1.2913 (±0.0105)
Musical Instruments	1.3505 (±0.0061)	4.0926 (±0.0164)	1.1151 (±0.0242)
Patio	1.2183 (±0.0096)	3.8782 (±0.0051)	1.1353 (±0.0174)
Automotive	1.0844 (±0.0084)	4.3252 (±0.0041)	1.0105 (±0.0186)
Average	1.2752 (±0.3729)	3.8122 (±0.5220)	1.1381 (±0.1020)

Data sets	ALFM	TransNet	deepLTRS
Fine Food	1.0705 (±0.0014)	1.3783 (±0.0012)	**0.9788** (±0.0215)
Musical Instruments	**0.8929** (±0.0013)	1.0912 (±0.0057)	0.9702 (±0.0143)
Patio	1.0219 (±0.0027)	1.0589 (±0.0009)	**0.9855** (±0.0319)
Automotive	**0.8797** (±0.0016)	1.0649 (±0.0012)	0.9299(±0.0511)
Average	0.9663 (±0.0819)	1.1483(±0.1334)	**0.9661** (±0.0392)

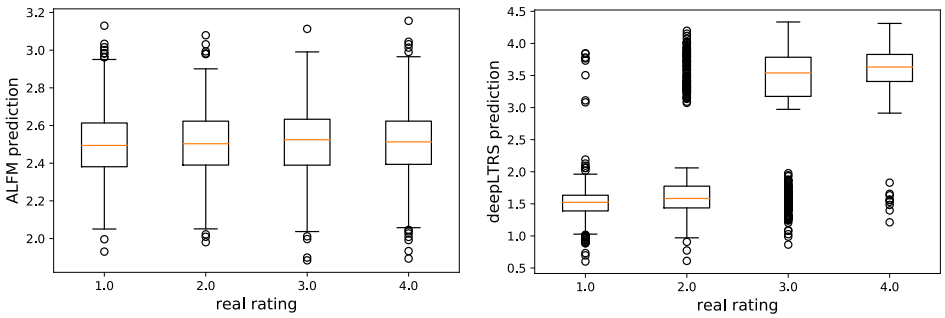


Figure 2: Comparisons of the predictions with actual ratings for ALFM and deepLTRS, where scores are simulated in the interval $[1, 4]$.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investment in the Future, project managed by the National Research Agency (ANR) with the reference numbers ANR-19-P3IA-0002.

Bibliographie

[Basbug and Engelhardt, 2017] Mehmet E Basbug and Barbara E Engelhardt. Coupled compound poisson factorization. arXiv preprint arXiv:1701.02058, 2017.
[Blei et al., 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
[Burke, 2002] Robin Burke. Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, 12(4):331–370, 2002.
[Catherine and Cohen, 2017] Rose Catherine and William Cohen. Transnets: Learning to transform for recommendation. In Proceedings of the eleventh ACM conference on recommender systems, pages 288–296, 2017.
[Cheng et al., 2018] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In Proceedings of the 2018 world wide web conference, pages 639–648, 2018.
[Dieng et al., 2019] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. arXiv preprint arXiv:1907.04907, 2019.
[Gopalan et al., 2015] Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with hierarchical poisson factorization. In UAI, pages 326–335, 2015.
[Herlocker et al., 2000] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 241–250, 2000.
[McAuley and Leskovec, 2013] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM conference on Recommender systems, pages 165–172, 2013.
[Mnih and Salakhutdinov, 2008] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In Advances in neural information processing systems, pages 1257–1264, 2008.
[Pazzani and Billsus, 2007] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In The adaptive web, pages 325–341. Springer, 2007.
[Srivastava and Sutton, 2017] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488, 2017.
[Zheng et al., 2017] Lei Zheng, Vahid Noroozi, and Philip S. Yu. Joint deep modeling of users and items using reviews for recommendation, 2017.

Heuristique de pente pour la régression linéaire en grande dimension

Perrine Lacroix[1,2,3] & Marie-Laure Martin-Magniette[1,2,4]

[1] *Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France.*
[2] *Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris Saclay (IPS2) 91405 Orsay, France*
[3] *Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, Université Paris-Saclay*
[4] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France*

perrine.lacroix@universite-paris-saclay.fr
marie_laure.martin-magniette@agroparistech.fr

Résumé

Nous considérons le problème de sélection de variables en régression linéaire gaussienne en grande dimension. Ce problème est motivé par des applications en génomique où l'acquisition de données d'expression est possible pour tous les gènes d'une entité simultanément. L'objectif est la reconstruction de réseaux de gènes ou l'identification des facteurs de transcription régulant l'expression d'un gène cible.
Nous proposons d'identifier les variables par minimisation d'un risque empirique pénalisé. La première étape de notre procédure est de régulariser la régression pour établir un ordre d'entrée des variables dans le support. Nous créons ainsi une collection de modèles dépendante du jeu de données. Lors de la seconde étape, nous proposons un nouveau critère non-asymptotique pour sélectionner le meilleur modèle pour un contrôle du risque prédictif. Il s'appuie sur l'heuristique de pente proposée par [Birgé and Massart, 2007]. La pénalité utilisée dépend de deux constantes à évaluer. Contrairement à la proposition de [Lebarbier, 2005], nous montrons que le rapport des deux constantes n'est pas fixe et qu'il est préférable de les calibrer toutes les deux. Pour cela, nous proposons un nouvel algorithme dont nous avons testé et comparé le comportement sur des simulations à partir de jeu de données différents.

Mots-clefs : Sélection de modèle, pénalisation, collection de modèle aléatoire, heuristique de pente.

Abstract

We study the problem of variable selection for a high-dimension Gaussian linear regression. This problem is motivated by genomics applications where gene expression measurements are available for all the genes simultaneously. The goal is either the reconstruction of the gene network or the identification of the transcription factors regulating the expression of one specific target gene.
We propose to identify the relevant variables by minimizing a penalized empirical risk. The first step of our procedure deals with the regularization of the regression to establish an order for the variable entry in the support. This step creates a data-dependent model collection. In the second step of our procedure, we propose a new non-asymptotic criterion to select the best model for a predictive risk control. It is based on the slope heuristic proposed by [Birgé and Massart, 2007]. The used penalty function depends on two unknown constants. Contrary to what [Lebarbier, 2005] proposed, we show that the ratio of the two constants is not constant and it is preferable to calibrate both. For that, we propose a new algorithm and its performances are tested and compared in a simulation study composed of different settings.

Keywords : Model selection, penalization, random model collection, slope heuristic.

1 Modèle

Nous considérons $(Y_i, X_{i1}, ..., X_{ip})_{i\in\{1,...,n\}}$ un échantillon de n observations indépendantes et identiquement distribuées selon le modèle statistique suivant :

$$Y = \sum_{j=1}^{p} \beta_j^0 X_j + \varepsilon, \qquad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Nous désignons par X la matrice de design fixe de taille $n \times p$ composée des vecteurs colonnes $(X_1, \cdots, X_p)$. Nous supposons que σ^2 est inconnue et que p est proche ou plus grand que n. L'objectif est d'estimer le p-vecteur $\beta^0 = \left(\beta_1^0, \cdots, \beta_p^0\right)$ pour assurer une bonne performance de prédiction.

2 Inférence

Pour pallier le problème de grande dimension, l'une des méthodes est d'ajouter une fonction de pénalité sur le risque empirique lors de sa minimisation afin de limiter le nombre de coefficients non nuls à estimer.

Une étude exhaustive de tous les sous-ensemble possible des p variables est impossible. Ainsi, la première étape consiste à restreindre l'estimation à une collection $\mathcal{M}$ de supports plausibles de β^0 donnant ainsi un ordre de pertinence sur les vecteurs colonnes $(X_1, \cdots, X_p)$. Nous supposons celle-ci réalisée par l'utilisation d'une pénalité telle que Lasso [Tibshirani, 1996] ou Elastic-Net [Zou and Hastie, 2005]. La seconde étape est basée sur la sélection du meilleur support parmi ceux de la collection.

Pour cette seconde étape, nous considérons l'approche proposée par [Birgé and Massart, 2001] qui repose sur la pénalisation des risques empiriques de la collection :

$$\text{crit}(m) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - (X\hat{\beta}_m)_i \right)^2 + \text{pen}(m) \right\}, \qquad \forall m \in \mathcal{M} \tag{2}$$

où pour chaque m désignant un modèle de la collection $\mathcal{M}$, S_m est un sous-espace de $\mathbb{R}^p$ de dimension D_m et $\hat{\beta}_m$ est l'estimateur tel que $X\hat{\beta}_m$ est la projection orthogonale de Y sur S_m. La pénalité est de la forme :

$$\text{pen}(m) = 2 \left(C_1(\sigma^2) \frac{D_m}{n} + C_2(\sigma^2) \frac{\log\left(\binom{p}{D_m}\right)}{n} \right). \tag{3}$$

L'enjeu est alors de trouver les constantes $C_1(\sigma^2)$ et $C_2(\sigma^2)$ pour minimiser correctement le risque quadratique. Généralement la pénalité est utilisée en fixant le rapport $\frac{C_1(\sigma^2)}{C_2(\sigma^2)}$ à 2.5, valeur obtenue par [Lebarbier, 2005] lors d'une large étude de simulation dans le cadre de la détection de rupture. La pénalité définie en (3) s'écrit alors

$$\text{pen}(m) = 2\kappa(\sigma^2) \left(2.5 \frac{D_m}{n} + \frac{\log\left(\binom{p}{D_m}\right)}{n} \right)$$

où $\kappa(\sigma^2)$ est calibrée par la méthode de l'heuristique de pente implémentée dans le package capushe du logiciel R [Baudry et al., 2012]. Cette calibration repose sur la relation linéaire attendue entre les valeurs du risque empirique et les valeurs $\left(2.5 \frac{D_m}{n} + \frac{\log\left(\binom{p}{D_m}\right)}{n} \right)$ pour les modèles m de grande dimension.

Ces travaux de [Birgé and Massart, 2007], [Lebarbier, 2005] et [Baudry et al., 2012] proposent donc une méthode d'estimation de la pénalité directement à partir des données permettant un ajustement plus fin de la

minimisation en fonction des données disponibles. Cependant, ils supposent une étude exhaustive possible de tous les modèles, éventuellement jusqu'à une certaine taille. Leur collection est alors dite fixe. Dans notre cadre, la collection est une sous-liste non exhaustive de modèles pertinents. Générée par un algorithme, elle est entièrement dépendante des données et donc aléatoire, ce qui peut altérer les performances d'une telle approche.

Nous proposons de tester la pénalité (3) lorsque la collection n'est plus fixe mais aléatoire et dépendante des données. Par des simulations, nous montrons qu'il est toujours possible d'observer un comportement linéaire entre les vecteurs $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}\right)$ et les valeurs du risque empirique pour les modèles de grande dimension. Cependant, nous constatons que le rapport $\frac{\hat{C}_1}{\hat{C}_2}$ n'est plus constant et nous proposons un nouvel algorithme pour calibrer les deux constantes C_1 et C_2.

3 Étude de simulation

Nous considérons plusieurs jeux de données simulés avec un bruit gaussien et des vecteurs colonnes X_j indépendants ou corrélés. Nous évaluons les performances de prédiction de notre procédure d'estimation et nous les comparons avec celles obtenues par (i) le critère eBIC [Chen and Chen, 2008] qui est le meilleur critère asymptotique pour la prédiction, (ii) l'heuristique de pente avec le rapport fixé à 2.5, (iii) le critère LinSelect [Giraud et al., 2012] qui permet un contrôle optimal non-asymptotique du risque prédictif.

4 Remerciements

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. IPS2 benefits from the support of the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS).

Références

[Baudry et al., 2012] Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics : overview and implementation. Statistics and Computing, 22(2) :455–470.

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). Gaussian model selection. Journal of the European Mathematical Society, 3(3) :203–268.

[Birgé and Massart, 2007] Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. Probability theory and related fields, 138(1-2) :33–73.

[Chen and Chen, 2008] Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95(3) :759–771.

[Giraud et al., 2012] Giraud, C., Huet, S., Verzelen, N., et al. (2012). High-dimensional regression with unknown variance. Statistical Science, 27(4) :500–518.

[Lebarbier, 2005] Lebarbier, É. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. Signal processing, 85(4) :717–736.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society : Series B (Methodological), 58(1) :267–288.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society : series B (statistical methodology), 67(2) :301–320.

A hidden semi-Markov model for segmenting environmental toroidal data

Francesco Lagona [1,3] & Antonello Maruotti [2,3]

[1] *University of Roma Tre, via Chiabrera 199 00145 Rome (Italy) - francesco.lagona@uniroma3.it*
[2] *LUMSA University, Via della Traspontina, 21 - 00193 Rome (Italy) - a.maruotti@lumsa.it*
[3] *Dept. of Mathematics, University of Bergen, Allégaten 41, 5007 Bergen (Norway)*

Abstract. Toroidal time series are temporal sequences of bivariate angular observations that often arise in environmental and ecological studies. A hidden semi-Markov model is proposed for segmenting these data according to a finite number of latent classes, associated toroidal densities. The model conveniently integrates circular correlation, multimodality and temporal auto-correlation. A computationally efficient EM algorithm is proposed for parameter estimation. The proposal is illustrated on a time series of wind and sea wave directions.

Keywords. hidden semi-Markov model, EM algorithm, model-based clustering, toroidal data

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. These data often arise in environmental and ecological studies. Examples include time series of wind and wave directions [8], time series of wind mean directions and directions of the maximum gust observed each day [2] and time series of turning angles in studies of animal movement [10].

The analysis of toroidal time series is complicated by the difficulties in modeling the dependence between angular measurements over time [7]. An additional complication is given by the multimodality of the marginal distribution of the data, because environmental toroidal data are observed under time-varying heterogeneous conditions.

This paper introduces a toroidal hidden semi-Markov model (HSMM) that simultaneously accounts for dependence across circular measurements, temporal auto-correlation, multimodality and latent time-varying heterogeneity. Under this model, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent semi-Markov process. While the toroidal density

accommodates dependence between two circular variables, a mixture of toroidal densities allows for multimodality and, finally, a latent semi-Markov process accounts for temporal correlation and, simultaneously, for time-varying heterogeneity.

Our proposal extends previous approaches that are based on toroidal hidden Markov models [9, 1]. Under a toroidal hidden Markov model, the data are approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent, first-order Markov chain with a finite number of states. The sojourn times of each state of a Markov chain are distributed according a geometric distribution. Hence the most likely dwell time for every state of a hidden Markov model with underlying first-order Markov chain is 1. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent semi-Markov model, allowing for sojourn times that are not necessarily geometrically distributed.

2 A hidden semi-Markov model for toroidal data

Let $\mathbf{z} = (x, y)$ be a pair of angles, $x, y \in [0, 2\pi)$. Moreover, let $f(x; \alpha)$ and $f(y; \beta)$ be two circular densities, respectively known up to the parameters α and β. Further, let $F(x; \alpha)$ and $F(y; \beta)$ be the two cumulative distribution functions of x and y, defined with respect to a fixed, although arbitrary, origin. Finally, let $g(u; \gamma), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z}; \theta) = 2\pi g\left(2\pi\left(F(x; \alpha) - qF(y; \beta)\right)\right) f(x; \alpha) f(y; \beta)) \quad q = \pm 1 \tag{1}$$

is a parametric toroidal density with support $[0, 2\pi)^2$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$ [3]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by decoupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then equation (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

The proposed hidden semi-Markov model can be described as a dynamic mixture of copula-based toroidal densities. To illustrate, let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t, y_t \in [0, 2\pi)$, be a toroidal time series. We assume that the distribution of the data is driven by the evolution of an unobserved semi-Markov process with K states, which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \ldots, K, \sum_k \pi_k = 1$, $K^2 - K$ transition probabilities $\pi_{hk} = P(u_{tk} =$

$1|u_{t-1,h} = 1), h, k = 1, \ldots, K, \sum_k \pi_{hk} = 1,\ h \neq k$ (whereas $\pi_{kk} = 0,\ k = 1 \ldots K$), and, finally, p parameters of the dwell time distributions of each state.

The specification of the HSMM is completed by assuming that the observations are conditionally independent, given a realization of the semi-Markov process. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{2}$$

where $f(\mathbf{z}; \theta_k)$, $k = 1, \ldots, K$ are the K cylindrical densities defined by (1) and known up to a vector of parameters θ_k.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{3}$$

By computing the maximum likelihood estimate $\hat{\theta}$ [11, chapter 12], the toroidal time series can be segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{4}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posteriori, MAP, allocation).

When the dwell distribution of each latent state is geometric, the model reduces to a hidden Markov model that ignores alternative dwell time distribution. If, additionally, the transition probability matrix of the model has equal rows, the model reduces to a mixture model where observations are clustered by ignoring the information redundancy that is due to temporal correlation.

3 An application to marine data

The proposed methods have been implemented to segment a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. Figure 1 displays the scatter plot of the data. Point coordinates indicate the direction (in radians) from which winds blow and waves travel. For simplicity, these bivariate observations are plotted on the plane, although data points are actually on a torus. The interpretation of these data is not easy. While in the ocean wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the orography of the basin and the

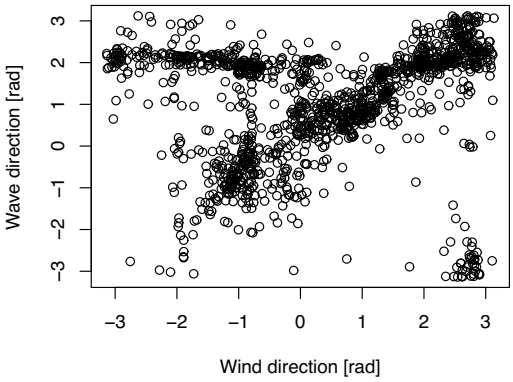


Figure 1: Wave directions and heights, as observed by the buoy of Ancona in wintertime ($-\pi$, $-pi/2$, 0, $\pi/2$ respectively indicate South, West, North, East). For simplicity, the data are plotted on the plane, although they are points on the torus $[-pi/2, \pi/2)^2$.

location of the buoy. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. When western and south-western winds blow from the coast, waves are not synchronized with wind and travel along the major axis of the Adriatic basin from SE to NW. This explains the clusters shown in Figure 1 and suggests the occurrence of two latent wind–wave regimes. Accordingly, a HSMM with two states have been estimated from these data.

The proposed HMM requires a parametric specification of the toroidal density (1), which reduces to the choice of the binding density g and the choice of the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the wind and wave direction. However, depending on the choice of the binding density, the density (1) can be multimodal [4]. Using multimodal densities in segmentation and classification problems, such as the one motivating this paper, may unnecessarily complicate the interpretation of the results. Unimodal densities can however be obtained by using the wrapped Cauchy as a binding density g [4].

Accordingly, for this study, the binding density has been specified as a centered wrapped Cauchy

$$g(u;\gamma) = \frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(u)} \quad u \in [0, 2\pi).$$

This circular density depends on a single concentration parameter $\gamma \in [0, 1)$ and reduces to the uniform circular density when $\gamma = 0$.

Wrapped Cauchy densities that include additional location parameters α_1 and β_1 have

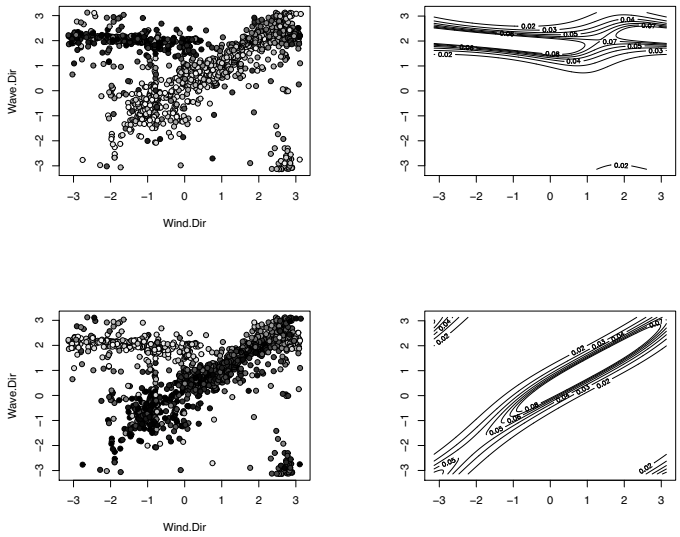


Figure 2: Segmentation of a time series of wind and wave directions. Left: observations colored with grey levels according to the estimated membership probabilities of each class (black indicates a probability equal to 1). Right: contour plot of state-specific toroidal densities.

been instead exploited to model the marginal distributions of wind and wave direction, say

$$f(x;\alpha) = \frac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(y-\alpha_1)} \quad x \in [0,2\pi) \tag{5}$$

$$f(y;\beta) = \frac{1}{2\pi}\frac{1-\beta_2^2}{1+\beta_2^2-2\beta_2\cos(y-\beta_1)} \quad y \in [0,2\pi) \tag{6}$$

The proposed toroidal density is therefore obtained by taking a wrapped Cauchy density that binds wrapped Cauchy marginals, a model known as the bivariate wrapped Cauchy model [5].

Figure 2 shows the shapes of the two state-specific toroidal distributions and the segmented observations. The model successfully segment the observations according to to clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, wind and wave directions are essentially independent, because coastal winds do not generate waves. Under state 2, winds blows along the major axis of the Adriatic basin and their directions are highly correlated with the directions of the wave that they generate.

Overall, the model describes the plasticity of the wind–wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime switching changes not only the modal directions and con-

centrations around these modes but also, and more interestingly, the correlation structure of the data. As a result, on the one side, the (marginal) weak correlation between wind and wave directions is explained by the presence of coastal winds (component 1). On the other side, the model indicates that the wind direction is an accurate predictor of the wave direction only under a specific regime (state 2).

References

[1] Bulla J, Lagona F, Maruotti A, Picone M (2012) A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series, Journal of Agricultural, Biological, and Environmental Statistics, 17: 544-567

[2] Coles S (1998) Inference for circular distributions and processes, Statistics and Computing, 8: 105-113.

[3] Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models. Journal of the American Statistical Association 73: 602-606.

[4] Jones MC, Pewsey A, Kato S (2015). On a class of circulas: copulas for circular distributions. Annals of the Institute of Statistical Mathematics 67: 843-862.

[5] Kato S, Pewsey A (2015) A Möbius transformation-induced distribution on the torus, Biometrika, 102: 359-370

[6] Lagona F (2019) Copula-based segmentation of cylindrical time series, Statistical and Probability Letters, 144: 16-22.

[7] Lagona F (2018) Correlated cylindrical data. In: C. Ley and T. Verdebout (Eds) Applied Directional Statistics: Modern Methods and Case Studies, Chapman & Hall/CRC: New York, 45-59.

[8] Lagona F, Picone M, Maruotti A, Cosoli S (2014) A hidden Markov approach to the analysis of space-time environmental data with linear and circular components, Stochastic Environmental Research and Risk Assessment 29: 397-409.

[9] Lagona F, Picone M (2013) Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data, Journal of Statistical Computation and Simulation, 83: 1223-1237

[10] Mastrantonio G (2018) The joint projected normal and skew-normal: A distribution for poly-cylindrical data, Journal of Multivariate Analysis, 165: 14-26.

[11] Zucchini W, Macdonald IL and Langrock R (2016) Hidden Markov models for time series, Chapman and Hall, Boca Raton FL (US)

Blind soil moisture inference

Sylvain LANNUZEL[1,2], Nicolas Gilardi[2], and Paul-Henry Cournède[1]

[1]Université Paris-Saclay, CentraleSupélec, Laboratoire de Mathématiques et Informatique, 91190 Gif-sur-Yvette, France
[2]CybeleTech, 92120, Montrouge, France

Abstract

L'humidité du sol est une variable clé dans la modélisation des cultures mais sa mesure est complexe et les capteurs adéquats sont souvent onéreux. Ce travail présente différentes méthodes permettant d'inférer la dynamique de l'humidité du sol à partir des seules données météorologiques.

Cette étude est préliminaire et a pour but d'expliquer la dynamique sur une station unique du réseau de capteurs Wegener Network.

Parmi les méthodes testées, les réseaux de neurones récurrents donnent des résultats très encourageants.

Soil moisture (SM) is a key variable in crop modelling, but its measurement can be hard, and appropriate recording devices are often costly. This work presents different methods in order to infer the soil moisture dynamics as a dynamic system with the sole climatic conditions as input data.

This is a preliminary study aiming at predicting the dynamic of a single recording SM station, from the Wegener Network.

Among the tested methods, recurrent neural networks give interesting results in reproducing the SM dynamic.

1 Introduction

1.1 Soil moisture

Soil moisture (SM) is the percentage of water in a unit volume of soil. It is a driving variable for outdoor agriculture as it has to be higher than the plants' water needs for it to have an optimal growth [5].

It can be seen as a dynamical system which can be written as :

$$y_{t+1} = f_\theta(y_t, u_t, u_{t-1}, \dots, u_{t-N}, p_t) \ , \tag{1}$$

where y_t is the soil moisture [%] at day t, θ describes the soil characteristics, u_t corresponds to the meteorological conditions on the same day and p_t is what is used by the plant (if any) for its growth.

SM is *fed* with the incoming water (from precipitation or irrigation) and consumed by the *evapotranspiration* [3], which is the combined effect of evaporation from the atmosphere and the use of water by the plant.

1.2 Soil characteristics

The soil characteristics θ express that different soils won't react the same way to similar weather conditions. The main parameter is the soil *texture* [5], which is determined by the quantity of *sand, silt and clay* in the soil. A *sandy* soil will have a poorer ability of retaining water than a *loamy* one which has a higher proportion of silt.

The soil is also described through its *field capacity*, which expresses the maximal quantity of water the soil can withhold, above which either the water is kept in surface (which can lead to flooding) or drained below the reach of the roots.

1.3 Meteorological variables

The process which rules the SM dynamic is described by the FAO in [3]. To use this model, meteorological variables are needed, including solar radiation (rad), mean temperature (TM), quantity of precipitation (prec) or air humidity (RH).

1.4 Dynamic aspect

SM value on day t is a function of the weather conditions described above as well as the SM values from the previous day(s). For instance, a wet and rainy day will tend to increase the water proportion in the soil, but this increase will be all the more pronounced as the quantity of water in the soil on the previous day is small. This rule of thumb can have more subtle effects, like the formation of a crust on the ground surface after several consecutive dry days [4].

1.5 Modelling

Empirical models [4, 7, 17] can reproduce accurately the SM dynamic but they often need a good calibration and suffer from poor generalization. To leverage this calibration issue, statistical approaches have been used to infer the SM value at a desired time horizon [6, 11, 15, 10, 2].

A paramount point of this study is to try to infer SM values **without** the knowledge of the previous ones during the prediction period (and thus without using expensive captors for this purpose). Only SM values during a learning phase are used. Similarly, we also suppose that no prior knowledge on the soil characteristics (pedological parameters) is available. The objective is to obtain an evaluation of this critical variable, with the sole knowledge of classical weather data, and thus with a far reduced cost.

2 Data description

2.1 Wegener Network

This work has been performed with data from a single station taken out of the Wegener Network [13] which provides a high-resolution dataset for different meteorological variables, including the ones needed by the FAO model, as well as the soil moisture recorded by a *Time Domain Reflectometer* (TDR), plotted on fig. 1.

Table 1 shows the correlation between the meteorological variables and SM as well as with the daily SM increment : $\Delta SM_t = SM_t - SM_{t-1}$. SM is negatively correlated with the mean temperature and the radiation, which can be explained as the soil is dryer in summer, where the radiation and temperature are higher. For the ΔSM, the highest correlation is with the precipitation, for this is the main input in the SM evolution process.

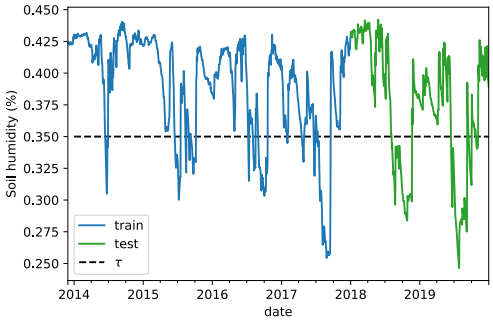


Figure 1: SM measurements on a single station of the Wegener Network. Data is split between a train set (blue) and a test set (green). The τ parameter (dashed line) indicates the threshold for the classification task

	TM	RH	prec	rad
SM	-0.34	0.11	0.03	-0.25
Δ SM	-0.05	0.23	0.44	-0.16

Table 1: Correlations between the SM signal and meteorological factors

The goal of this report is to evaluate the different methods enabling to infer the daily SM value based on meteorological variables, but also to detect days when SM is below a threshold, chosen to indicate situations with water shortage for plant growth *i.e.* when $y_t < \tau$, here set to 0.35.

3 Soil moisture inference

In this section we will review the different methods for evaluating SM on a single station based on meteorological data. We will first present the mechanistic model, and then statistical methods. Inferred value at time t will be written as $\hat{y}_t$. Two

This station has been recording data with soil moisture since late 2013, and was therefore split between a training set (2013-2017) and a test set (2018-2019) fig. 1.

3.1 Mechanistic model

The STICS model [4] is built of blocks which each describes a subprocess of plant growth.

Among those subprocesses, the water balance calculates the available water to the plant on a given day. This subprocess is a balance between what is incoming to the system (rain and irrigation) and what is lost through evaporatranspiration or drained below the reach of the roots.

This model depends on parameters θ which will rule the dynamic of this process. An acceptable range of values for these parameters is given in [4] but they have to be precisely estimated in order to fit the soil characteristics under study. For simplicity, we use here a simple least-square estimate. Other frequentist or Bayesian methods can bed used in such configuration [18, 8], notably for a better assessment of parameter uncertainty, or to take into account available prior information.

Let $\mathcal{M}_\theta$ be the subprocess evaluating SM based on the daily meteorological data. To find the best parameters we seek to solve :

$$\arg\min_\theta (\mathcal{M}_\theta(U) - y)^2 \tag{2}$$

Where U is the time series corresponding to the meteorological data (precipitation, temperature, relative humidity, radiation and wind speed) and y is the recorded SM by the TDR probe.

To begin with, we chose initial values for each parameter and evaluated the evolution of the reconstruction error with a single parameter varying and the other ones fixed. From the second plot we can straightforwardly see that the optimization problem eq. (2) is non convex. Note that the y-axis on this second plot has been changed for readability purpose as the MSE variation's range was of the order of 10^{-1}. To solve this non convex problem, we used the Powell method [14] which does not need gradient and works for bound-constrained optimization problems fig. 2.

The result is shown on fig. 3. Some parameters have estimated values which match very closely the upper or lower bound of the constraint intervals (corresponding to the acceptable range of values given by [4]), but broadening those bounds or choosing different initial positions led to the same conclusion. The mechanistic model appears to underestimate SM during dry episodes, thus exaggerating the evaluation of water stress for plants. We can imagine the consequences if decision aid tools were based on this model: irrigation could be recommended inappropriately. Moreover, the estimated output is noisier than the original data.

*

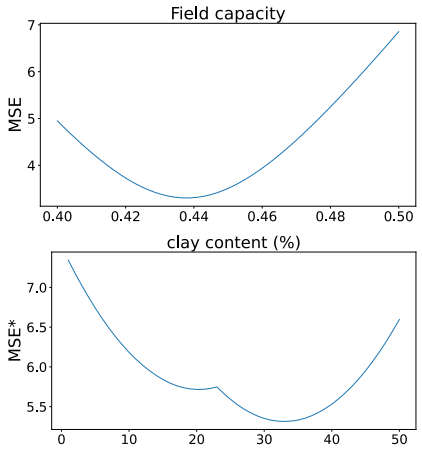


Figure 2: Mono-dimensional variation of $(\mathcal{M}_\theta(U) - y)^2$ for two different parameters of the STICS model eq. (2)

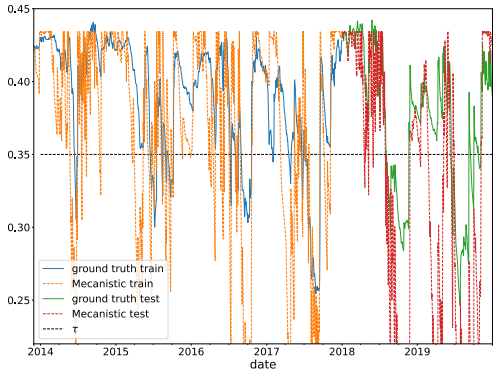


Figure 3: Optimized STICS reconstruction on the reference station

3.2 Statistical models

Table 1 shows that relationships exist between the meteorological variables and SM. Statistical models can then be used to try to mimic these dependencies and enable inference based on new data.

3.2.1 Regression model

The first model we considered is a simple linear regression. We seek to find θ such that $||X^T\theta - y||^2$ is as small as possible, where X is a feature matrix.

Two scenarios are used to build X, one (**S1**) that uses only the previous meteorological data as input : $X_t = (u_t, \dots, u_{t-l})$ and one that combines an auto-regressive part (**S2**) and is built such that

- During training : $X_t = (u_t, y_{t-1}, \dots, u_{t-l}, y_{t-l-1})$
- During evaluation : $X_t = (u_t, \hat{y}_{t-1}, \dots, u_{t-l}, \hat{y}_{t-l-1})$

This second linear model enables to carry the temporal information from the meteorological variables as well as SM. This also abides by our will to have a *blind* approach, as the inference on test data will be done without the knowledge of the true SM data. l has been chosen through a cross-validation procedure to 5 days.

Results for both configurations are plotted on figs. 4 and 5

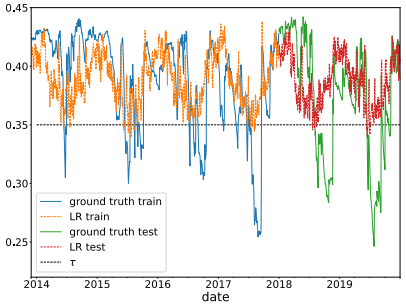


Figure 4: Linear regression result **S1**

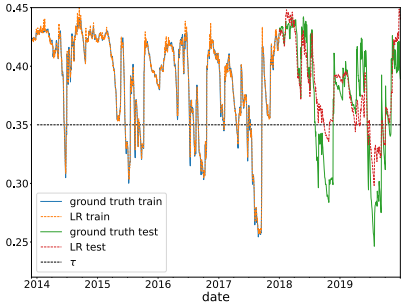


Figure 5: Linear regression results **S2**

*y-axis is changed for readability purpose on clay content plot

There is not enough information in the sole weather variables (**S1**) to catch the SM dynamic, only the seasonal trend is explained by this model. **S2** however shows that an estimation of the previous SM values combined with the weather data are sufficient to evaluate very accurately the daily SM. This result is expressed by the almost perfect fit of the estimated and ground truth data for the training set. However, the error appears when using the inferred value of the previous day as a proxy for the last SM value, as can be seen on the test set. The model tends to overestimate the low SM values and using inferred value as input can be a source of drift with time even though such behaviour is not observed here. The impact is inverse to the one discussed for the mechanistic models: a water stress episode for the crops could be missed.

3.2.2 LSTM

Results from section 3.2.1 tend to indicate that linear models are too basic to reproduce the SM dynamic. However, the very positive impact of introducing the autoregressive part in the liner model inspired us to use Recurrent Neural Networks (RNN). They are special cases of neural networks applied to time series. An intuitive diagram of such a network is presented on fig. 6. The idea is to describe the system's state through a *state variable* h_t. This state variable is computed from the so-called the *encoder* **E** :

$$h_t = \mathbf{E}(X_t, h_{t-1})$$

Ideally, in our scenario, this state variable is supposed to capture information relevant to the SM dynamic and value. Once this state variable is calculated, the other part of the network, the *decoder* **D** will translate this information into the desired SM evaluation.

$$\hat{y}_t = \mathbf{D}(h_t)$$

There are various available architectures for these two networks. For regression purposes, the decoder is usually a simple linear regression :

$$\hat{y}_t = h_t^T \theta_{\mathbf{D}}$$

For the training procedure, examples are fed to the network, by creating batches of n consecutive days $((X_{t_1}, y_{t_1}), \ldots, (X_{t_n}, y_{t_n}))$, which produce SM evaluations at each time step. A *backpropagation* [16] algorithm is run by re-adjusting the weights of the network based on the error between SM evaluation and ground truth.

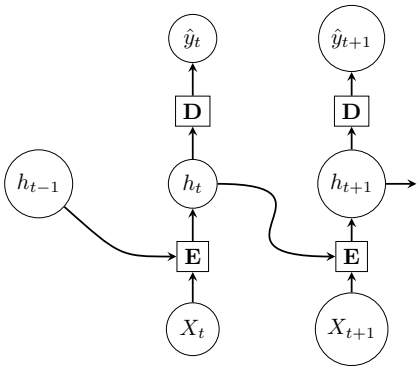


Figure 6: Schematic diagram of a Recurrent Neural Network

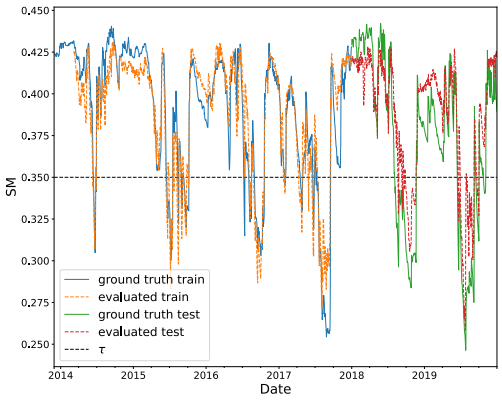


Figure 7: LSTM reconstruction on the reference station

The state variable in the network captures information about the dynamic of the system, but this temporal effect can be quickly attenuated depending on the encoder type. This happens for dense RNN, and more sophisticated networks have been built to overcome this issue, such as Long Short Term Memory networks (LSTM) [12].

To choose the most appropriate architecture for the evaluated problem, we performed a grid-search procedure whose parameters were the encoder type, the state size and the feature length. We used TensorFlow [1] with the Keras [9] framework. The grid-search result showed that the most appropriate network was an LSTM with a state size of 2 and training length of 100 days.

The evaluation of this trained model on the reference station data is presented on fig. 7.

3.3 Comparison results

The global result is shown in table 2. Models were evaluate by comparing inferred SM value to ground truth value on both training and test data. We used Root Mean Squared Error (RMSE) as a metric to quantify their regression capacities and $F1$-score for the classification sub-problem. Bold elements represent the best value for each metric on test data. The best model is the LSTM even though, a regression with an auto-regressive part has also good predictive power.

Table 2: Comparison results for different inference models

model	LR (S1)		LR (S2)		LSTM		Mechanistic	
	train	test	train	test	train	test	train	test
RMSE	3.48	4.52	0.44	2.60	1.77	**2.40**	4.99	9.15
F1 score	0.13	0.09	0.97	0.71	0.87	**0.83**	0.59	0.75

4 Conclusion

This preliminary study shows the potential of statistical methods such as simple linear models or more elaborate recurrent neural networks.

This work will be extended with other statistical approaches, including a more extensive study of auto-regressive methods using exogenous variables. Once all methods are set, they will be evaluated on more than a single recording station. The Wegener Network is composed of 12 stations which record different SM dynamics for multiple soil types, this is a challenge when it comes to creating a single model for different soil characteristics. We will also investigate other open datasets to take full advantage of data greedy methods such as LSTM.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, and G. Research. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Technical report, Google research, 2015.

[2] K. O. Achieng. Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models. *Computers & Geosciences*, 133, 2019.

[3] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Technical report, FAO, 1998.

[4] N. Brisson, C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoche, D. Zimmer, J. Sierra, P. Bertuzzi, P. Burger, F. Bussière, Y. M. Cabidoche, P. Cellier, P. Debaeke, J. P. Gaudillère, C. Hénault, F. Maraux, B. Seguin, and H. Sinoquet. An overview of the crop model STICS. In *European Journal of Agronomy*, volume 18, pages 309–332. Elsevier, 1 2003.

[5] C. Brouwer, A. Goffeau, and M. Heibloem. *Irrigation Water Management: Training Manual No. 1-Introduction to Irrigation*. FAO, 1985.

[6] Y. Cai, W. Zheng, X. Zhang, L. Zhangzhong, and X. Xue. Research on soil moisture prediction model based on deep learning. *PLoS ONE*, 14(4), 4 2019.

[7] A. Chanzy, M. Mumen, and G. Richard. Accuracy of top soil moisture simulation using a mechanistic model with limited soil characterization. *Water Resources Research*, 44(3), 3 2008.

[8] Y. Chen and P. H. Cournède. Data assimilation to reduce uncertainty of crop model prediction with-Convolution Particle Filterin. *Ecological Modelling*, 290(C):165–177, 2014.

[9] F. Chollet and Others. Keras, 2015.

[10] W. Dorigo, W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, E. Haas, P. D. Hamer, M. Hirschi, J. Ikonen, R. De Jeu, R. Kidd, W. Lahoz, Y. Y. Liu, D. Miralles, T. Mistelbauer, N. Nicolai-Shaw, R. Parinussa, C. Pratola, C. Reimer, R. Van Der Schalie, S. I. Seneviratne, T. Smolander, and P. Lecomte. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Biogeochemistry Remote Sensing of Environment*, 203:185–215, 2017.

[11] M. K. Gill, T. Asefa, M. W. Kemblowski, and M. McKee. SOIL MOISTURE PREDICTION USING SUPPORT VECTOR MACHINES. *Journal of the American Water Resources Association*, 42(4):1033–1046, 8 2006.

[12] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[13] G. Kirchengast, T. Kabas, A. Leuprecht, C. Bichler, and H. Truhetz. WegenerNet: A Pioneering High-Resolution Network for Monitoring Weather and Climate. *Bulletin of the American Meteorological Society*, 95(2):227–242, 2 2014.

[14] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1 1964.

[15] N. J. Rodríguez-Fernández, F. Aires, P. Richaume, Y. H. Kerr, C. Prigent, J. Kolassa, F. Cabot, C. Jiménez, A. Mahmoodi, and M. Drusch. Soil moisture retrieval using neural networks: Application to SMOS. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):5991–6007, 11 2015.

[16] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[17] P. Steduto, T. C. Hsiao, D. Raes, and E. Fereres. Aquacrop-the FAO crop model to simulate yield response to water: I. concepts and underlying principles. *Agronomy Journal*, 101(3):426–437, 5 2009.

[18] S. Trevezas and P. H. Cournède. A Sequential Monte Carlo Approach for MLE in a Plant Growth Model. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2):250–270, 6 2013.

Estimation of multivariate generalized gamma convolutions through Laguerre expansions

O. Laverny [1,2,*] & E. Masiello [1] & V. Maume-Deschamps [1] & D. Rullière [3]

[1] *Institut Camille Jordan, UMR 5208, Université Claude Bernard Lyon 1, Lyon, France*

[2] *SCOR SE*

[3] *Mines Saint-Étienne, UMR CNRS 6158, LIMOS, Saint-Étienne, France*

[*] *Corresponding Author, laverny@univ-lyon1.fr*

Résumé. La classe des convolutions généralisées de lois gammas fut créée par Thorin dans le but d'établir la divisibilité des lois log-Normales et Pareto. Bien que ces distributions fussent étudiées de manière extensive dans le cas univarié, le cas multivarié et les structures de dépendance qui en découlent ont reçu peu d'intérêt dans la littérature. De plus, aucune procédure d'estimation pour ces distributions n'est connue. En exprimant les densités des convolutions généralisés de loi gamma multivariés dans une base de Laguerre tensorisée, nous comblons le manque et fournissons des procédures d'estimation pour les cas univariés et multivariés. En étudiant la performance de ces procédures, nous fournissons également une expansion en série convergente pour les densités de convolution de lois gammas multivariés, que nous montrons comme étant plus stable que les séries univariés de Moschopoulos et Mathai. Nous présentons finalement quelques exemples.

Mots-clés. Convolutions généralisées de lois gammas multivariées, GGC, mesure de Thorin, base de Laguerre, infinie divisibilité, estimation. ...

Abstract. The generalized gamma convolutions class of distributions appeared in Thorin's work while looking for the infinite divisibility of the log-normal and Pareto distributions. Although these distributions have been extensively studied in the univariate case, the multivariate case and the dependence structures that can arise from it have received little interest in the literature. Furthermore, no estimation procedures are available for these distributions. By expanding the densities of multivariate generalized gamma convolutions into a tensorised Laguerre basis, we bridge the gap and provide estimation procedures for both the univariate and multivariate cases. We provide some insight about performance of these procedures, and a convergent series for the density of multivariate gamma convolutions, which is shown to be more stable than Moschopoulos's and Mathai's univariate series. We furthermore present some examples.

Keywords. Multivariate Generalized Gamma Convolutions, GGC, Thorin's measure, Laguerre's basis, infinite divisibility estimation. ...

1 Introduction to Generalized Gamma Convolutions

The class $\mathcal{G}_1$ of univariate generalized gamma convolutions, defined as weak limits of independent convolutions of gamma distributions, was first introduced by Thorin (1977) as a tool to show the infinite divisibility of log-normal and Pareto distributions. By definition, $\mathcal{G}_1$ is closed under independent convolutions, but as Bondesson (2015) recently showed it is also closed by independent products of random variables. Pareto, log-Normal, α-stable, Weibull, and many other useful distributions belong to this class, which makes it a nice framework for many application fields such as climate events modeling, insurance, etc.

Analogous multivariate classes $\mathcal{G}_{d,n}$ and $\mathcal{G}_d$, $d \geq 1$, were constructed by Bondesson (2009), following an old idea of Cherian (1941). The class $\mathcal{G}_{d,n}$ contains convolutions of n comonotonic random vectors with gamma marginals: a random vector $\boldsymbol{X} \in \mathcal{G}_{d,n}$ follows the additive risk-factor structure:

$$\begin{aligned} X_1 =& Y_{1,1} + \ldots + Y_{1,n} \\ & \ldots \\ X_d =& Y_{d,1} + \ldots + Y_{d,n}, \end{aligned} \tag{1}$$

where $Y_{i,j}$ are all gamma distributed, each row vector $Y_{i,\cdot}$ has independent marginals, and each column vector $Y_{\cdot,j}$ has comonotonous marginals with common shape. Then $\mathcal{G}_d$ is constructed as the closure of all $\mathcal{G}_{d,n}$ classes. Since on one hand some $Y_{i,j}$ might be identically zero, as $0 \in \mathcal{G}_{1,1}$, and on the other hand every $Y_{i,j}$ is infinite divisible, by increasing n the model can achieve a wide variety of dependence structures, and marginal distributions can approach any distribution in $\mathcal{G}_1$. Last but not least, the dependence structure is naturally asymmetric and can have a wide range of shapes, including tail dependence or independence.

Using an ad hoc Laguerre basis to express densities of these distributions, we provide procedures for parameters estimation, which was not possible at the current stage of the literature. We provide some examples of our estimator.

2 Thorin's measure and Miles's projection

Let $\boldsymbol{X} \in \mathcal{G}_d$. Then there exists a measure ν, the so-called Thorin measure, such that the cumulant generating function K of X can be expressed as:

$$K(\boldsymbol{t}) = \ln\left(\mathbb{E}\left[e^{\langle \boldsymbol{t}, \boldsymbol{X}\rangle}\right]\right) = -\int_{R_+^d} \ln\left(1 - \langle \boldsymbol{t}, \boldsymbol{s}\rangle\right) \nu\left(\partial \boldsymbol{s}\right).$$

When $\boldsymbol{X} \in \mathcal{G}_{d,n}$, ν is n-atomic and atoms and weights of ν are respectively scales and shapes of gamma vectors $Y_{\cdot,j}$ from (1). Therefore, the model parameters can be fully summarized by ν, which we would like to estimate from data.

The current literature contains no estimation procedure for distributions in $\mathcal{G}_d$ or $\mathcal{G}_{d,n}$, but only one projection procedures from known densities, given by Miles, Furman & Kuznetsov (2019), which finds an approximation in $\mathcal{G}_{1,n}$ to a formal density in $\mathcal{G}_1$.

Their result is based on the fact that, when $d = 1$, the first derivative of K is a Stieltjes function which has specific useful properties when the initial density is exactly a density in $\mathcal{G}_1$. Furthermore, they require the evaluation of the moment generating function $M(\boldsymbol{t}) = \exp\{K(\boldsymbol{t})\}$ derivatives at $\boldsymbol{t} = -\boldsymbol{1}$ (exponentially shifted moments) with 300 digits precision, excluding e.g. estimation from empirical datasets. From these moments, they frame the problem of estimation of ν into a generalized moment problem, which can be efficiently solved since K' is a univariate Stieltjes function. Sadly, if the initial information does not strictly correspond to a $\mathcal{G}_d$ density, this moment problem frequently does not have a solution at all.

Here, we investigate the estimation of d-variate Thorin measures through a least-square approach to this moment problem. We now focus on the subclass $\mathcal{G}_{d,n}$ [1].

3 A loss through a certain Laguerre basis

The set of d-variate Laguerre functions exposed by Dussap (2020),

$$\forall \boldsymbol{k} \in \mathbb{N}^d, \quad \phi_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{i=1}^{d} \phi_{k_i}(x_i) \text{ where } \phi_k(x) = \sqrt{2} e^{-x} \sum_{p=0}^{k} \binom{p}{k} \frac{(-2x)^p}{p!},$$

[1] A continuous Thorin measure is not easy to work with: there is no simulation procedure for the distribution, no known expressions for the density or the distribution function, and even the Laplace transform requires numerical integration for evaluation.

constitute an orthonormal basis of the set of square-integrable functions $L_2(\mathbb{R}_+^d)$. The coefficients of a $\mathcal{G}_{d,n}$ density in this basis are a certain linear combination of the (exponentially shifted) moments, i.e. the derivatives at $\boldsymbol{t} = -\mathbf{1}$ of the moments generating function $M(\boldsymbol{t}) = \exp\{K(\boldsymbol{t})\}$. Fortunately, the derivatives of K are easy to obtain for $\mathcal{G}_{d,n}$ models, and through a fast recursive version of Faà di Bruno formula given by Miatto (2019) we are able to compute Laguerre coefficients efficiently.

From this expansion, we obtain an approximation of an Integrated Square error loss between densities from the parameters of the model. Let $\boldsymbol{X}_1, ...\boldsymbol{X}_N$ be a random sample from a density f with expansion $f = \sum_{\boldsymbol{k}\in\mathbb{N}^d} a_{\boldsymbol{k}}\phi_{\boldsymbol{k}}$. First, the $\boldsymbol{k}$-th Laguerre coefficients $a_{\boldsymbol{k}}$ can be estimated by Monte-Carlo:

$$\widehat{a_{\boldsymbol{k}}} = \frac{1}{N}\sum_{i=1}^{N} \phi_{\boldsymbol{k}}(\boldsymbol{X}_i).$$

Then, if we denote coefficients of a $\mathcal{G}_{d,n}(\boldsymbol{\alpha}, \boldsymbol{s})$ distribution by $a_{\boldsymbol{k}}(\boldsymbol{\alpha}, \boldsymbol{s})$, we estimate the parameters $\boldsymbol{\alpha}, \boldsymbol{s}$ of our model from random samples by the following optimization program:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{s}}) = \underset{\boldsymbol{\alpha}\geq 0, \boldsymbol{s}\geq 0}{\arg\min} \sum_{\boldsymbol{k}\leq \boldsymbol{m}} \left(\widehat{a_{\boldsymbol{k}}} - a_{\boldsymbol{k}}(\boldsymbol{\alpha}, \boldsymbol{s})\right)^2,$$

where $\boldsymbol{m}$ is a threshold to control the size of the basis.

The remaining part of the integrated square error (that would have an expression in the $\boldsymbol{k} > \boldsymbol{m}$ part of the orthogonal basis) can be shown to go to zero as the number of observations and the size of the basis increase.

An analysis of this loss show that it is constructed as a very high degree polynomial in the parameters $\boldsymbol{\alpha}, \boldsymbol{s}, M(-\mathbf{1})$. It has therefore a myriad of local minima, and a global optimizer should be used. We found that Particle Swarms routines from Zhan, Zhang, Li, & Chung (2009) work quite well for these problems.

Several outcomes can be highlighted:

- Through an error bound on the Laguerre coefficients inside $\mathcal{G}_{d,n}$, we provide a bound on the error of the series expansion for densities in $\mathcal{G}_{d,n}$, where the current literature gives densities relying on Mathai and Moschopoulos series, which work only when $d = 1$ and are known to be unstable or even dramatically failing for certain parameters ranges, noteworthy those which correspond to projection of log-Normal, Pareto or Weibull distributions.

- The estimation algorithm still works for densities that are not in the class, e.g. a Weibull with a shape of $\frac{3}{2}$ which is projected, with an irreducible error, onto the class.
- The algorithm is essentially the same whatever the dimension of the initial data, and can use standard empirical data in 64-bits precision as a target.

Comparisons with Mile's projection in a univariate setting show very good results. As the multivariate case has no other estimation procedures in the literature, we propose to illustrate our results with graphs of the dependence structure and marginals obtained through estimation of a convolution of 20 bivariate gammas, taken on a simulation from a Clayton copula with log-Normal and Pareto marginals, in Figure 1.

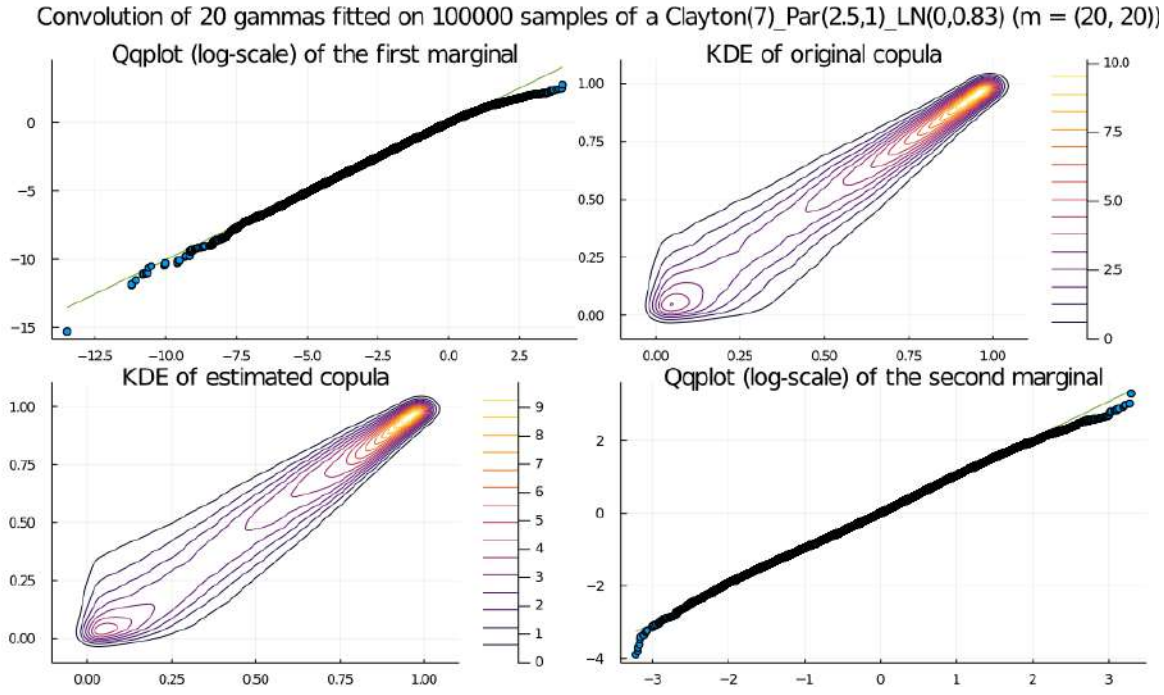


Figure 1: Estimation of a $\mathcal{G}_{2,20}$ distribution from $N = 100000$ simulations of a Clayton($\theta = 7$) copula with Pareto($\alpha = 2.5$) and log-Normal($\mu = 0, \sigma = 0.83$) marginals. We present two quantile-quantile plots for the marginals, and Gaussian kernels of 100000 pseudo-observations from the original sample and from a sample of the estimated distribution.

We remark that the dependence structure was estimated quite correctly. The problems in the marginal upper tails could be easily overcome by increasing the number of gammas, as both the log-Normal and the Pareto distributions are in $\mathcal{G}_1$.

Bibliographie

Bondesson, L. (2015), A class of probability distributions that is closed with respect to addition as well as multiplication of independent random variables, *Journal of Theoretical Probability*, 28, 1063–1081

Bondesson, L. (2009), On univariate and bivariate generalized gamma convolutions, *Journal of Statistical Planning and Inference*, 139, 3759–3765

Cherian, K.C. (1941), A bivariate correlated gamma-type distribution function, *Journal of the Indian Mathematical Society*, 5, 133–144

Dussap, F. (2020), Anisotropic multivariate deconvolution using projection on the Laguerre basis

Miatto, F.M. (2019), Recursive multivariate derivatives of $e^{f(X_1,...,X_n)}$ of arbitrary order, *arXiv:1911.11722*

Miles, J. & Furman, E. & Kuznetsov, A. (2019), Risk aggregation: a general approach via the class of generalized gamma convolutions, *Variance*, in press

Thorin, O. (1977), On the infinite divisibility of the Pareto distribution, *Scandinavian Actuarial Journal*, 1977, 31–40

Zhan, Z. H., Zhang, J., Li, Y., & Chung, H. S. H. (2009), Adaptive particle swarm optimization, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), 39(6), 1362-1381

Estimation des paramètres d'un modèle de culture à partir de données de plein champ et de données de plateforme de phénotypage

Jean-Benoist Leger [1] & Estelle Kuhn [2] & Boris Parent [3] & François Tardieu [4] Claude Welcker [5]

[1] *UTC, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France, estelle.kuhn@inrae.fr*
[3] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, boris.parent@inrae.fr*
[4] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, francois.tardieu@inrae.fr*
[5] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, claude.welcker@inrae.fr*

Résumé. Les modèles de culture élaborés par des écophysiologistes décrivent les processus de développement d'une plante. Ils permettent en particulier de rendre compte des différences de comportement de plusieurs variétés dans différents environnements, dues aux interactions génotype-environnement. Pour les utiliser à des fins prédictives, il est nécessaire de calibrer auparavant leurs paramètres. Nous considérons le modèle de culture APSIM et proposons un modèle joint bayésien à effets mixtes dans lequel nous inférons la valeur des paramètres inconnus à partir de données issues d'expérience de plein champ et mesurées en plateforme de phénotypage. Nous choisissons des lois *a priori* informatives pour intégrer les connaissances d'expert et implémentons un algorithme de type Gibbs hybride pour simuler la loi *a posteriori*. Les résultats obtenus sur données simulées et réelles mettent en évidence le gain obtenu sur la précision des estimations en utilisant les données issues de plateforme de phénotypage en sus des données du champ.

Mots-clés. Modèle de culture, données hétérogènes, modèles à effets mixtes, modèle bayésien, algorithme Gibbs hybride.

Abstract. Crop models were developed by ecophysiologists to describe plant development. They allow in particular to report difference existing between several genotypes in several environments, due to genotype by environment interaction. It is first necessary to calibrate these models to use them for prediction purpose. We consider the crop model APSIM and present a joint bayesian model with mixed effects. We infer models parameter values from data collected in the field and in phenotyping platform. Prior distribution are chosen in order to integrate expert knowledge. We implement an hybrid Gibbs algorithm to simulate the posterior distribution. Results obtained from simulated and real data highlight clearly the advantage of using phenotyping platform data in addition to field data.

Keywords. crop model, heterogeneous data, mixed effects models, bayesian model, Gibbs hybrid algorithm.

1 Introduction

1.1 Contexte de l'amélioration des plantes

Un des enjeux actuels en sciences du végétal vise à mieux comprendre les mécanismes impliqués dans le développement des plantes et leurs réponses aux conditions environnementales. Ces mécanismes diffèrent d'une espèce à l'autre, chacune ayant des phases de développement spécifiques. Au sein d'une même espèce, les différents processus qui se succèdent au cours de la croissance ont lieu de façon différente selon la variété considérée : ils se réalisent plus ou moins rapidement, à des périodes plus ou moins précoces, donnant lieu à une importante variabilité de comportements. Mieux comprendre comment cette variabilité dans les processus de croissance est reliée au génotype caractérisant la variété est un objectif essentiel en amélioration des plantes.

De plus, une forte interaction existe entre la variété considérée et l'environnement, incluant non seulement les aspects météorologiques mais également la composition du sol, les intrants, etc. Ainsi, une même variété va évoluer différemment dans différents environnements et différentes variétés vont se comporter différemment dans un même environnement (cf. Millet et al. (2016)). Mieux comprendre ces interactions génotype-environnement-conduite de culture est un levier important pour fournir de meilleures recommandations dans le choix des variétés selon l'environnement, en particulier dans un contexte de changement climatique fort, incluant de plus en plus d'événements climatiques extrêmes (cf. Millet et al. (2019)).

La modélisation mathématique est un outil extrêmement pertinent pour mieux comprendre, quantifier et prédire ces interactions. Des modèles linéaires ont tout d'abord été utilisés, donnant lieu à un cadre relativement limité du point de vue de la modélisation des effets génotypiques et environnementaux. Plus récemment, des modèles descriptifs des mécanismes de croissance des plantes, appelés modèles de culture, ont été développés par des écophysiologistes des plantes, comme par exemple le modèle APSIM (cf. Keating et al (2003)). Ces modèles dynamiques rendent compte des processus qui interviennent lors du développement de la plante. Ils utilisent en entrée des variables environnementales et des paramètres dépendant du génotype et fournissent en sortie des caractères plus ou moins intégrés de la plante comme par exemple la date de floraison, le rendement, la biomasse au cours du temps. Ces modèles descriptifs permettent également de modéliser les interactions génotype-environnement-conduite de culture. Utiliser à des fins prédictives, ils sont un outil efficace pour prédire ces interactions. Cependant un grand nombre de paramètres de ces modèles sont généralement inconnus et doivent être calibrés. La valeur des paramètres peut être ajustée manuellement en comparant les sorties du modèle à des données. Cette approche requiert néanmoins beaucoup de temps, d'autant plus si le nombre de paramètres est important. Une approche plus rapide consiste à ajuster un modèle statistique basé sur le modèle de culture à partir des données disponibles, en inférant la valeur des paramètres via un estimateur statistique. Ce type d'approche reste néanmoins

difficile à mettre en oeuvre du fait de la complexité du modèle de culture et demande la mise en place de méthodes numériques efficaces. Des premières approches ont été proposées (cf. Cooper et al (2016)). Cependant elles ne permettent de traiter qu'un nombre réduit de données et d'estimer un petit nombre de paramètres.

1.2 Le modèle de culture APSIM

Nous considérons le modèle de culture Agricultural Production Systems sIMulator (APSIM) avec le module maïs et une extension du module feuille réalisée par l'unité INRAE LEPSE (Lacube et al. (2017)). Il s'agit d'un modèle à pas de temps journalier qui simule un couvert constitué de plantes moyennes en mettant à jour un vecteur de descripteurs de la plante qui évolue au cours du temps. Les entrées de ce modèle sont des covariables météorologiques, des covariables descriptives du sol et de la conduite de culture. Il fournit en sortie la date de floraison et les composantes du rendement. Certains paramètres ont un sens physique, comme par exemple l'efficacité d'interception lumineuse, d'autres ne sont pas interprétables. Par ailleurs, certains processus sont communs à toute l'espèce, leurs paramètres ont une valeur commune à tous les génotypes de cette espèce, tandis que d'autres processus sont variables génétiquement, leurs paramètres ont des valeurs différentes selon le génotype, comme par exemple le nombre final de feuilles du plant.

1.3 Les données de sources hétérogènes

Nous disposons d'un riche jeu de données issu du projet DROPS European Project incluant un panel de diversité composé de 230 génotypes hybrides observés en plein champ dans 13 conditions environnementales. Une condition environnementale est définie par un lieu et une année. Les lieux sont répartis en Europe du nord au sud, rendant compte de conditions climatiques très contrastées. Les années considérées varient de 2012 à 2013. Les données comprennent la date de floraison, les composantes du rendement (nombre de grains, poids d'un grain) et les conditions environnementales.

De plus, des expériences auxiliaires complémentaires ont été réalisées sur la plateforme INRAE PhénoArch à Montpellier (Cabrera-Bosquet et al. (2016)). Ces expériences ont permis d'obtenir des données supplémentaires sur des paramètres mécanistes intervenant dans le modèle de culture. Ainsi, des mesures de quantités telles que le nombre final de feuilles observées ont été effectuées en plateforme, apportant des informations complémentaires au jeu de données obtenu en plein champ.

L'objectif est d'utiliser les deux sources d'information champ et plateforme dans la procédure d'inférence statistique des paramètres du modèle de culture.

2 Modélisation statistique

2.1 Modélisation de la variabilité génotypique

Nous disposons pour chaque génotype du panel DROPS de mesures répétées du rendement, du nombre de grains et de la date de floraison dans M conditions environnementales. Nous considérons un modèle statistique à effets mixtes (cf. Pinheiro et Bates (2000)) basé sur le modèle de culture APSIM qui permet de prendre en compte simultanément les variabilités inter-génotype et intra-génotype. On note Y_{gm} la mesure vectorielle dans la condition expérimentale m pour le génotype g et on modélise pour tout g et tout m :

$$\log Y_{gm} = \log G(e_m, \beta_g, \gamma_g) + \varepsilon_{gm} \tag{1}$$

où G représente la sortie du modèle de culture APSIM, e_m le vecteur de variables descriptives de l'environnement m, β_g le vecteur des paramètres du génotype g non mesurées en plateforme, γ_g le vecteur des paramètres du génotype g mesurées en plateforme, ε_{gm} un terme d'erreur supposé gaussien centré de variance diagonale $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. On suppose que les vecteurs d'effets aléatoires (β_g) et (γ_g) sont indépendants identiquement distribués de loi gaussienne d'espérance $\bar{\beta}$ resp. $\bar{\gamma}$, et de matrice de covariance diagonale Σ_β, resp. Σ_γ.

2.2 Modélisation des données de plateforme de phénotypage

Nous considérons un modèle joint pour intégrer les données issues de la plateforme à l'inférence des paramètres du modèle de culture. Pour cela, nous modélisons les mesures des paramètres effectuées en plateforme de phénotypage via un modèle linéaire en fonction de la valeur des paramètres du modèle de culture APSIM. On note Z_g le vecteur de taille p des mesures de paramètres du génotype g. Pour $1 \leq l \leq p$, on a :

$$Z_{g,l} = \mu_l + \zeta_l \gamma_{g,l} + \eta_{g,l} \tag{2}$$

où μ et ζ sont des vecteurs inconnus de $\mathbb{R}^p$ et $\eta_{g,l}$ un terme d'erreur résiduel supposé gaussien centré de variance τ_l^2.

3 Inférence bayésienne du modèle joint

Du fait de la complexité du modèle de culture APSIM et du grand nombre de paramètres à estimer, nous faisons le choix d'une approche bayésienne qui va permettre de régulariser la procédure d'estimation. Nous souhaitons choisir des lois *a priori* uniformes pour les paramètres mécanistes du modèle de culture qui prennent leurs valeurs dans des intervalles bornés fixés suivant des dires d'expert. Toutefois, pour des raisons computationnelles, nous avons effectué une reparamétrisation du modèle, et les nouveaux paramètres sont à valeurs réelles. Nous choisissons pour ces paramètres des lois *a priori* normales, telles que leurs transformées par la reparamétrisation inverse soient les plus proches au sens de

la divergence de Kullback-Leibler des lois uniformes de départ. Pour les paramètres μ et ζ du modèle des données issues de la plateforme, nous choisissons des lois *a priori* normales centrées sur la valeur attendue, 0 pour l'ordonnée à l'origine et 1 pour la pente. Nous fixons des lois inverse gamma qui sont conjuguées pour les lois *a priori* des paramètres de variances des bruits.

Nous appliquons un algorithme de Monte Carlo Markov Chain de type Gibbs hybride (cf. Carlin et Louis (2008)) pour générer une chaine de Markov qui sous des hypothèses de régularité du modèle est ergodique et a pour loi stationnaire la loi *a posteriori*. A partir des réalisations de cet algorithme, nous construisons des estimateurs empiriques des lois *a posteriori* des paramètres du modèle.

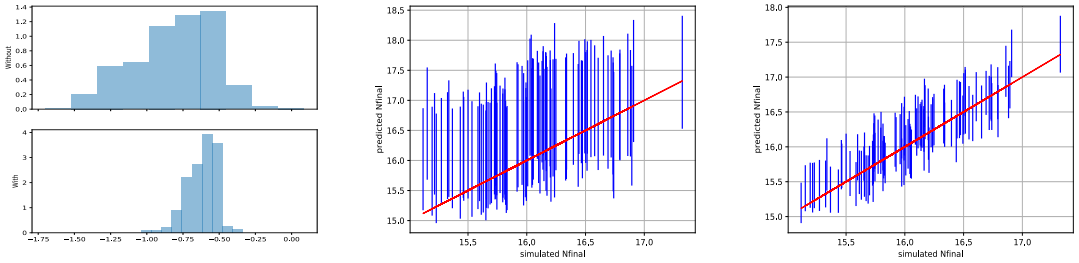


Figure 1: Histogramme de la loi *a posteriori* de Nfinal sans (gauche, haut) et avec (gauche, bas) les données plateforme supplémentaires ; Simulations versus prédictions via intervalles de crédibilité à 90% pour Nfinal sans (centre) et avec (droite) les données plateforme supplémentaires.

4 Expériences numériques

Nous effectuons une étude de simulation en considérant les 13 environnements réels du jeu de données DROPS et les valeurs des paramètres proches de ceux du génotype de référence *B*73. Nous simulons 100 génotypes. Nous estimons les trois paramètres du modèle correspondants au nombre final de feuilles (noté Nfinal), au premier ligulochrone et au poids moyen potentiel d'un grain, les autres étant fixés à la valeur de référence. Nous mettons en évidence que les estimateurs obtenus à partir des données issues du champ et de la plateforme dans le modèle joint sont plus précis que les estimateurs obtenus à partir des seules données issues du champ dans le modèle initial (cf Figures 1 et 2 gauche).

Nous ajustons ensuite le modèle proposé aux données réelles. Les prédictions obtenues à partir du modèle avec les paramètres estimés à partir des données issues du champ et de la plateforme sont meilleures que celles obtenues avec les paramètres estimés à partir des seules données champ (cf Figure 2 droite).

Ce travail a été financé par le projet AMAIZING ANR-10-BTBR-01. Les auteurs remercient la plateforme MIGALE, INRAE, 2020, Migale bioinformatics Facility pour les moyens de calcul et les capacités de stockage.

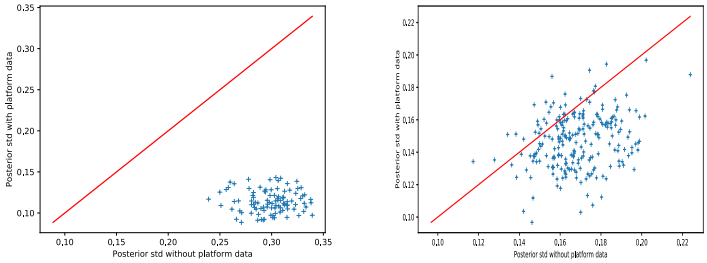


Figure 2: Ecart-types de la distribution *a posteriori* de Nfinal obtenus sans (abscisse) et avec (ordonnée) les données plateforme supplémentaires en simulation (gauche) et sur données réelles (droite).

Bibliographie

Cabrera-Bosquet, L., et al., (2016), High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, 212, (1), 269-281.

Carlin, B.P. and Louis, T. (2008), Bayesian methods for data analysis, *Chapman and Hall/CRC*.

Cooper, M. and Technow, F. and Messina, C. and Gho, C and Totir, L. R. (2016), Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial, *Crop Science*, 56, (5), 2141–2156.

Keating, B. and Carberry, P. and Hammer, G. and Probert, M. and Robertson, M. and Holzworth, D. and Huth, N.and Hargreaves, J. and *et al.*, (2003), An overview of APSIM, a model designed for farming systems simulation, *European journal of agronomy*, 18, (3-4), 267–288.

Lacube, S., et al., (2017) Distinct controls of leaf widening and elongation by light and evaporative demand in maize, *Plant Cell and Environment*, 40, (9), 2017-2028.

Millet E., Welcker C, Kruijer W, Negro S, Coupel-Ledru A, et al., (2016), Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios, *Plant Physiol*, 172, 749-764.

Millet, E. and Kruijer, W. and Coupel-Ledru, A. and Prado, S.A. and Cabrera-Bosquet, L. and Lacube, S. and Charcosset, A. and Welcker, C. and van Eeuwijk, F. and Tardieu, F., (2019), Genomic prediction of maize yield across European environmental conditions, *Nature genetics*, 51, (6), 952–956.

Pinheiro, J.C. and Bates D.M. (2000), Mixed-Effects Models in S and S-PLUS, *Springer*.

Estimateur de l'usage des codons dans le translatome

Carine Legrand [1] & Francesca Tuorto [2]

[1] *Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany, c.legrand@dkfz.de and Independent researcher, Kreuzstr. 5, 68259 Mannheim, Germany, carine.legrand1@gmail.com*
[2] *Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany, f.tuorto@dkfz.de*

Résumé. Les méthodes de profilage ribosomique donnent accès aux segments d'ARN messager protégés par le ribosome, pendant la traduction. Le séquençage de ces segments fournit la fréquence d'usage des codons sur l'ensemble des ARN messagers, le translatome. La normalisation et l'estimation de cette fréquence sont peu examinés et ne font pas consensus. Nous présentons un nouvel estimateur pour les coefficients d'usage des codons. Nous examinons ses caractéristiques en termes de biais et de convergence et présentons une application sur une lignée de cellules tumorales.

Mots-clés. ARN, traduction, profilage ribosomique

Abstract. Ribosome profiling allows to obtain ribosome-protected mRNA fragments during translation. Sequencing of these fragments yields codon usage over the pool of mRNA, namely the translatome. The normalisation and estimation methods to calculate this codon usage are rarely examined and there is no consensus on the best method. We present a new estimator of codon usage. We investigate its bias and convergence and show an application on simulated data and in a cancer cell line.

Keywords. RNA, translation, ribosome profiling

1 Introduction

La traduction de l'ARN messager en protéines est un phénomène clé, finement régulé par les cellules d'un organisme, et qui se trouve en équilibre entre la transcription du génome d'une part, et la régulation du protéome d'autre part. Le profilage ribosomique permet de mesurer directement le translatome, ensemble des ARN messagers étant activement traduits par des ribosomes (Ingolia et al. 2009). Cette méthode repose sur l'arrêt et l'isolation des ribosomes, puis sur l'extraction des brins d'ARNm présents dans le ribosome, et finalement leur séquençage. Souvent, le profilage ribosomique est utilisé pour déterminer la fréquence d'utilisation des codons et notamment l'usage différencié de codons synonymes. Cependant, peu de travaux se sont penchés sur les caractéristiques

des estimateurs de ces quantités. Nous proposons un nouvel estimateur et examinons ses caractéristiques. Nous présentons une application à une lignée de cellules tumorales en présence ou non du nutriment queuine (Tuorto et al. 2018, Legrand et Tuorto 2020).

2 Contexte et estimateur proposé

La plupart des estimateurs utilisent la normalisation du nombre de codons dans le site accepteur du ribosome par le nombre total de codons, et par la fréquence n constatée sur les codons voisins, par exemple :

$$\tilde{E}_{c,A} = \frac{\frac{n_{c,A}}{\sum\limits_{c \in codons} n_{c,A}}}{\frac{1}{3}\left(\frac{n_{c,-3}}{\sum\limits_{c \in codons} n_{c,-3}} + \frac{n_{c,-2}}{\sum\limits_{c \in codons} n_{c,-2}} + \frac{n_{c,-1}}{\sum\limits_{c \in codons} n_{c,-1}}\right)}, \tag{1}$$

où c désigne un codon parmi l'ensemble des codons {aaa, aac, ...}, A désigne l'emplacement dans le site accepteur du ribosome, -3 désigne le $3^{ème}$ codon en amont du site A, etc. L'inconvénient de cette approche est que l'on prend pour hypothèse que le ribosome n'interagit pas avec ces codons voisins, ce qui est peu réaliste puisque ces codons sont également protégés par le ribosome. Hussmann (2015) a proposé un estimateur plus élaboré, tenant compte de la fréquence des codons observée sur tout le translatome, qui fournit une mesure à toute position autour du site A, cependant cette mesure est biaisée.

L'estimateur de l'usage du codon d'identité c que nous proposons est le suivant (Legrand et Tuorto 2020) :

$$\tilde{E}_{c,i} = \bar{E}_{c,i} = \frac{\sum_g n_{c,i,g}}{\sum_{c,g} n_{c,i,g} \cdot codon\, usage_c^{global}}, \tag{2}$$

où g désigne un ARNm et l'usage global des codons est défini par :

$$codon\, usage_c^{global} = \frac{\sum_g (\sum_c n_{c,0,g}) \cdot codon\, usage_{c,g}}{\sum_{c,g} n_{c,0,g}}. \tag{3}$$

L'estimateur (2) converge vers l'unité, lorsqu'un codon est utilisé de façon neutre (ni ralenti ni accéléré), et en posant l'hypothèse que l'usage des codons ne dépend pas de leur position sur l'ARNm.

3 Application

Cet estimateur a été appliqué à la lignée de cellules tumorales HeLa en présence ou non de queuine (Tuorto et al. 2018). La queuine est une base fréquemment substituée à la

guanine dans le site anticodon de certains ARN transfert. Cette base n'est pas synthétisée par l'organisme et doit être apportée par l'alimentation. La figure 1 montre l'estimateur de Hussmann, peu bruité mais biaisé, et l'estimateur proposé.

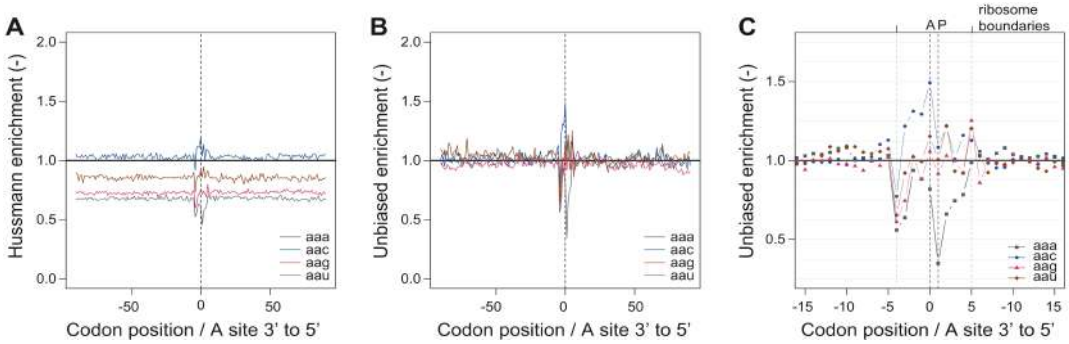


Figure 1: Estimateur de Hussmann (A) et estimateur proposé (B,C)

Cet estimateur permet par ailleurs de déterminer l'usage des codons pour chaque échantillon, en plus de permettre les comparaisons entre différentes conditions. Des précisions seront apportées, notamment sur la mise au point de l'estimateur, l'examen de son biais et sa convergence, lors de la présentation orale.

Bibliographie

Hussmann,J.A., Patchett,S., Johnson,A., Sawyer,S. and Press,W.H. (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. PLoS Genetics, 11, e1005732.

Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, pp. 218223.

Legrand,C. and Tuorto,F. (2020) RiboVIEW: a computational framework for visualization, quality control and statistical analysis of ribosome profiling data. Nucleic Acids Research, 48, e7plos g.

Tuorto,F., Legrand,C., Cirzi,C., Federico,G., Liebers,R., Muller,M., Ehrenhofer-Murray,A.E., Dittmar,G., Grone,H.J. and Lyko,F. (2018) Queuosine-modified tRNAs confer nutritional control of protein translation. EMBO J., 37, e99777.

Robustesse dans le modèle des blocs latents : application au test de positionnement en langues SELF

Margaux Leroy [1] & Vincent Brault[1] & Sylvain Coulange[2] & Marie-Pierre Jouannaud[3] & Frédérique Letué[1] & Marie-José Martinez[1] & Anne-Cécile Perret[4]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France.*
[2] *Univ. Grenoble Alpes, LIDILEM, 38000 Grenoble, France.*
[3] *Univ. Grenoble Alpes, LE, 38000 Grenoble, France.*
[4] *Univ. Grenoble Alpes, CUEF, 38000 Grenoble, France.*
prenom.nom@univ-grenoble-alpes.fr

Résumé. Dans cet exposé, nous nous intéressons à l'estimation des paramètres d'un modèle des blocs latents dans le cadre d'un test de positionnement en langue SELF et à une procédure de sélection de modèle afin de créer des groupes homogènes d'étudiants mais aussi des groupes homogènes de questions du test. Nous étudions ensuite la robustesse de cette procédure à partir d'une étude de simulation menée sur un jeu de données réelles. Nous nous intéressons à deux indicateurs : le nombre de groupes d'étudiants sélectionné et l'appartenance de deux étudiants donnés à un même groupe en fonction de la taille de l'échantillon.

Mots-clés. Modèles des blocs latents, robustesse

Abstract. In this presentation, we address the parameter estimation issue in the latent block models (LBM) framework applied to a language placement test called SELF, and we consider a model selection procedure to build homogeneous groups of students but also homogeneous groups of test questions. We next study the robustness of this procedure through a simulation study from a real data set. Two indicators are considered : the selected number of students groups and the membership of two given students to a same group with respect to sample size.

Keywords. Latent Block Model, robustness

1 Introduction

Les étudiants entrant à l'université en France, de niveaux hétérogènes en langues, ont besoin d'être évalués, puis dirigés vers des groupes de langues de niveaux différents pour leur permettre de progresser au mieux. Les universités disposent pour cela de différents tests de positionnement. Le test SELF, développé à Grenoble, est l'un des plus utilisés. A

∗. Institute of Engineering Univ. Grenoble Alpes

l'issue du test, chaque étudiant se voit attribuer un score agrégé, qui correspond au niveau du cours dans lequel il doit s'inscrire, et un niveau pour chacune des trois compétences évaluées (compréhension de l'oral, de l'écrit et expression écrite).

Par ailleurs, les concepteurs des tests ont de leur côté besoin d'évaluer si les questions qu'ils proposent sont pertinentes pour l'évaluation, en particulier, si elles sont suffisamment discriminantes. Il peut alors s'avérer utile de constituer des groupes de questions plus ou moins difficiles par compétence.

Les résultats d'un test peuvent se présenter sous la forme d'une matrice où une ligne correspond à un étudiant et une colonne à une question. L'élément (i, j) de la matrice vaut 1 si le $i^{\text{ème}}$ étudiant a réussi la $j^{\text{ème}}$ question, et 0 sinon. Les modèles des blocs latents s'avèrent alors particulièrement utiles pour constituer des groupes homogènes d'étudiants et de questions.

Les algorithmes usuels pour estimer les paramètres dans un modèle des blocs latents (*Variational* proposé par Govaert et Nadif (2008) ou *Stochastic Expectation Maximization* proposé par Keribin et al (2010)) sont dérivés de l'algorithme *EM*.

Une difficulté majeure dans les modèles des blocs latents est de sélectionner correctement le nombre de groupes d'étudiants et le nombre de groupes de questions. Deux critères principaux émergent de la littérature : le *Bayesian Information Criterion* (BIC) et le critère *Integrated Completed Likelihood* (ICL) . Notre procédure est basée sur le critère ICL et sur la maximisation de l'énergie libre (voir Keribin et al. (2015)).

Le but de l'exposé est d'étudier la robustesse de la procédure proposée. La robustesse est ici définie en termes de stabilité des groupes d'étudiants quand le nombre d'étudiants varie.

2 Modélisation et estimation

Dans cette partie, nous présentons le modèle des blocs latents et l'estimation de ses différents paramètres.

Fixons le nombre d'étudiants à n, le nombre de groupes d'étudiants à g, le nombre de questions à q et le nombre de groupes de questions à m. On appelle $Z_i, i = 1 \cdots n$, les variables aléatoires indépendantes modélisant le groupe d'étudiants auquel appartient l'étudiant i de loi multinomiale $\mathcal{M}(1; \pi_1, \ldots, \pi_g)$, et $W_j, j = 1 \cdots q$, les variables aléatoires indépendantes modélisant le groupe des questions auquel appartient la question j de loi multinomiale $\mathcal{M}(1; \rho_1, \ldots, \rho_m)$.

Une fois les groupes d'étudiants et de questions fixés, on suppose les réponses aux questions indépendantes. Sachant que le $i-$ème étudiant appartient au groupe k et la $j-$ème question appartient au groupe l, on modélise la réponse Y_{ij} par une loi de Bernoulli de paramètre α_{kl} :

$$P(Y_{ij} = 1 | Z_i = k, W_j = l) = \alpha_{kl}.$$

Pour un nombre de groupes d'étudiants fixé g et un nombre de groupes de questions

fixé m, les paramètres $\pi_1, \ldots, \pi_g, \rho_1, \ldots, \rho_m, \alpha_{11}, \ldots, \alpha_{gm}$ sont estimés via l'algorithme *V-Bayes* pour éviter le problème des groupes vides induit par l'algorithme VEM, combiné avec un échantillonneur de Gibbs pour limiter le problème des valeurs initiales (voir Keribin et al. (2015)).

3 Procédure de sélection de modèle

Pour choisir de manière optimale les nombres de groupes g et m, une solution serait de parcourir une grille de couples (k, l) et de sélectionner le couple qui maximise le critère *Integrated Completed Likelihood* (ICL) (Keribin et al. (2015)). Or le coût de calcul d'une telle procédure serait trop important. Nous passons donc par l'intermédiaire de l'énergie libre, qui est asymptotiquement équivalente à l'ICL mais possède un coût de calcul raisonnable. Notre procédure consiste donc, pour un couple (k, l) fixé, à

1. maximiser l'énergie libre ;
2. obtenir le couple (z, w) associé ;
3. calculer le critère ICL associé à ce couple (z, w).

On obtient ainsi un critère ICL par couple (k, l) de la grille. Le couple optimal (g, m) est celui qui maximise les ICL sur la grille.

4 Robustesse de la procédure

Nous nous intéressons ici à la robustesse de la procédure proposée dans les sens suivants :

— le nombre de groupes d'étudiants en fonction de la taille de l'échantillon,
— l'appartenance de deux étudiants donnés à un même groupe en fonction de la taille de l'échantillon.

Pour cela, nous nous appuyons sur un jeu de données réelles issu du test de positionnement SELF en anglais constitué de 228 étudiants ayant répondu à un test de 36 questions.

Dans un premier temps, nous appliquons la procédure décrite ci-dessus au jeu de données entier en faisant varier les nombres possibles de groupes (g, m) de 1 à 10 chacun. Le critère ICL nous permet de sélectionner 3 groupes d'étudiants et 5 groupes de questions qui nous servent de référence par la suite.

Pour une taille d'échantillon d'étudiants fixée ($n = 20, 60, \ldots, 220$), nous tirons 100 fois un échantillon d'étudiants de taille n sans remise parmi les 228 étudiants de départ. Nous appliquons la procédure à ces 100 échantillons. Dans la partie 5, nous présentons, pour chaque taille d'échantillon n, les distributions du nombre de groupes d'étudiants et du nombre de groupes de questions sélectionnés par le critère ICL.

Nous comparons ensuite la partition obtenue des n étudiants avec la partition de référence. Lorsque le nombre de groupes d'étudiants sélectionné est égal au nombre de groupes d'étudiants de référence ($g = 3$), nous calculons les effectifs communs à chaque groupe pour chaque permutation possible des labels des groupes en utilisant la version unidimensionnelle du critère utilisé par Lomet et al. (2012). Lorsque le nombre de groupes n'est pas le même, nous testons toutes les réunions possibles de groupes ; dans ce cas, une faible erreur de classification signifie que les groupes obtenus sont des réunions des groupes de référence (si la valeur estimée de g est égale à 2) ou des subdivisions (si la valeur estimée de g est supérieure ou égale à 4).

5 Résultats

Dans le Tableau 1, nous représentons la distribution des couples (g, m) sélectionnés par la procédure décrite précédemment. Nous pouvons observer que, plus le nombre d'étudiants augmente (colonne de gauche sur le Tableau 1), plus les nombres (g, m) se concentrent autour de 3 à 4 groupes d'étudiants et de 4 à 5 groupes de questions (colonne de droite sur le Tableau 1). Rappelons que le couple (g, m) de référence est $(3, 5)$ (symbolisé par un carré dans le tableau).

TABLE 1 – Distribution des couples (g, m) sélectionnés par la procédure proposée en fonction de n. Le carré symbolise le couple de référence.

$n = 20$

g \ m	2	3	4	5	6	7	Total
2	52	5	0	0	0	0	57
3	29	7	1	0	0	0	37
4	6	0	0	0	0	0	6
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

$n = 140$

g \ m	2	3	4	5	6	7	Total
2	0	1	7	2	0	0	10
3	0	9	30	17	1	0	57
4	0	2	22	5	2	1	32
5	0	0	0	0	0	0	0
6	0	0	0	0	1	0	1

$n = 60$

g \ m	2	3	4	5	6	7	Total
2	1	28	0	0	0	0	29
3	0	41	12	1	0	0	54
4	0	11	2	2	0	0	15
5	0	1	1	0	0	0	2
6	0	0	0	0	0	0	0

$n = 180$

g \ m	2	3	4	5	6	7	Total
2	0	0	3	0	0	0	3
3	0	0	31	22	3	0	56
4	0	0	21	11	4	0	36
5	0	0	0	2	0	1	3
6	0	0	0	1	1	0	2

$n = 100$

g \ m	2	3	4	5	6	7	Total
2	0	10	7	0	0	0	17
3	0	23	28	6	1	0	58
4	0	13	9	1	1	0	24
5	0	0	0	1	0	0	1
6	0	0	0	0	0	0	0

$n = 220$

g \ m	2	3	4	5	6	7	Total
2	0	0	0	0	0	0	0
3	0	0	17	31	2	0	50
4	0	0	14	11	11	1	37
5	0	0	0	6	7	0	13
6	0	0	0	0	0	0	0

Sur la Figure 1, nous présentons pour chaque taille d'échantillon n la répartition du pourcentage d'étudiants mal classés en fonction du nombre de groupes d'étudiants sélectionné. Remarquons que ce pourcentage ne peut pas dépasser $(g-1)/g$ (Robert et al. (2020)). Nous observons que, plus la taille d'échantillon n augmente, plus le pourcentage d'étudiants mal classés diminue quand on a sélectionné le bon nombre de groupes $g = 3$ (boxplots verts). Cependant, sur les échantillons où le nombre de groupes d'étudiants sélectionné est égal à 4 (boxplots bleus), ce pourcentage reste constant montrant que le quatrième groupe n'est pas simplement une subdivision d'un des groupes de référence.

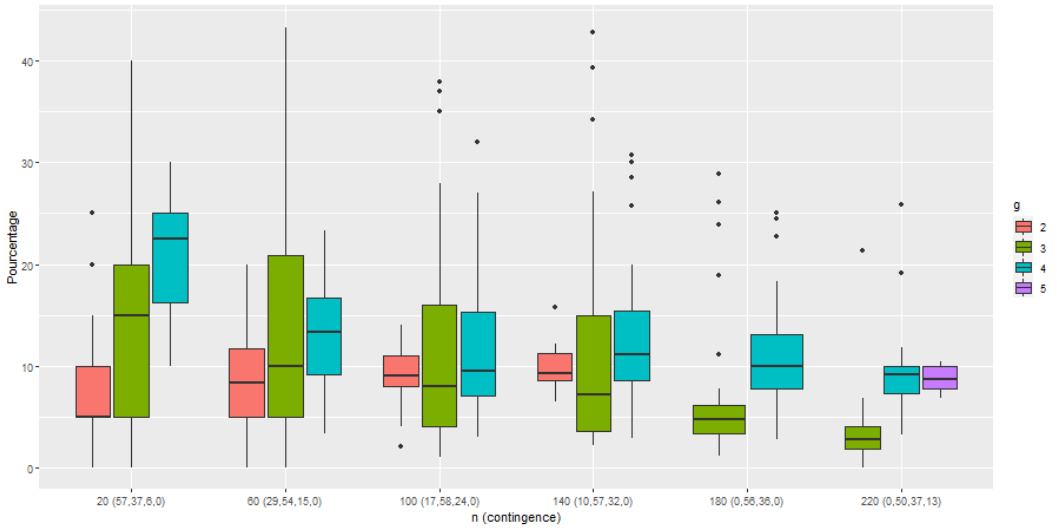


FIGURE 1 – Répartition du pourcentage d'étudiants mal classés en fonction de la taille d'échantillon n et du nombre de groupes sélectionné par la procédure (couleurs des boxplots). Seuls les boxplots concernant au moins 5 individus ont été conservés ; le nombre d'échantillons par boxplot est indiqué entre parenthèses à la suite de n.

Bibliographie

Bhatia, P., Iovleff, S., et Govaert, G. (2014). Blockcluster : an R package for model based co-clustering.

Biernacki, C., Celeux, G., et Govaert, G. (2000). Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719-725.

Govaert, G., et Nadif, M. (2008). Block clustering with Bernoulli mixture models : comparison of different approaches. *Computational Statistics and Data Analysis*, 52, 3233-3245.

Keribin, C., Govaert, G., et Celeux, G. (2010). Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*.

Keribin, C., Brault, V., Celeux, G., et Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6), 1201-1216.
Lomet, A., Govaert, G., et Grandvalet, Y. (2012). Design of artificial data tables for co-clustering analysis. *Université de Technologie de Compiègne*, France.
Robert, V., Vasseur, Y., et Brault, V. (2020). Comparing high-dimensional partitions with the Co-clustering Adjusted Rand Index. *Journal of Classification*, 1-29.

Inférence statistique pour un processus de dégradation en présence de maintenances : le modèle ARD 1

Margaux Leroy [1]

[1] *Université Grenoble Alpes, Laboratoire Jean Kuntzmann, margaux.leroy@univ-grenoble-alpes.fr*

Résumé. On s'intéresse ici à des systèmes industriels ou technologiques qui se dégradent au cours du temps. Ces systèmes sont soumis à des actions de maintenance dont l'effet est de réduire le niveau de dégradation. Afin de modéliser ce type de dégradation, plusieurs modèles ont déjà été proposés dans la littérature. On considèrera ici le modèle ARD 1 (Arithmetic Reduction of Degradation) et on s'intéressera à l'estimation de ses paramètres. On distinguera quatre cas de figure selon la manière dont les niveaux de dégradation sont observés. Dans le premier cas, on considère qu'on observe les niveaux de dégradation juste avant et juste après chacune des maintenances. Dans le second (resp. troisième) cas, on observe le niveau de dégradation juste avant (resp. après) la maintenance, mais pas juste après (resp. avant). Enfin dans le quatrième cas, on n'observe ni le niveau de dégradation juste avant ni le niveau de dégradation juste après les maintenances. Chacun de ces quatre cas amène à une écriture différente de la vraisemblance des observations, et donc à des estimations différentes.

Mots-clés. Modèles de dégradation, Maintenances imparfaites, Estimation paramétrique

Abstract. In this article, we consider technological or industrial equipments that are subject to degradation. These systems undergo maintenance actions, which reduce the degradation level. Several models have already been proposed for this type of degradation. Here we focus on the ARD 1 model (Arithmetic Reduction of Degradation). We study the estimation of its parameters. We will give four different approaches according to the way the degradation levels are observed. In the first case, we assume that we observe the degradation levels just before and after each maintenance. In the second (resp. third) case, we observe the degradation level just before (resp. after) maintenance but not after (resp. before). In the last case, we do not observe the degradation level neither before nor after the maintenances. Each case leads to a different writing of the likelihood of the observations and therefore to different estimations.

Keywords. Deterioration modeling, Imperfect maintenance models, Parametric Estimation

1 Présentation du modèle ARD 1 : *Arithmertic Reduction of Degradation*

Afin de modéliser un processus de dégradation au cours du temps, on décide de recourir à des processus stochastiques de Wiener. Les accroissements d'un tel processus sont indépendants et de loi normale. Soit X un processus de Wiener, on écrira $X(t) = \mu\, t + \sigma\, B(t)$ où B est un mouvement Brownien. On a $X(t+\Delta t) - X(t) \sim \mathcal{N}(\mu\Delta t\ ,\ \sigma^2\Delta t)$.

A certains instants donnés, on décide d'effectuer des maintenances dont l'effet est de réduire le niveau de dégradation. Le modèle ARD 1 propose une modélisation de cet effet des maintenances.

Le modèle ARD 1 a été proposé par Mercier-Castro [1]. L'hypothèse de ce modèle est que l'effet des maintenances est de réduire le niveau de dégradation d'une quantité proportionnelle au niveau de dégradation accumulé depuis la dernière maintenance.
Le facteur de proportionnalité ρ est le paramètre d'efficacité de la maintenance. Celui-ci vaut 0 lorsque les maintenances sont sans effet et vaut 1 lorsqu'elles sont parfaites (elles remettent le système à neuf). La valeur de ce paramètre est identique d'une maintenance à l'autre pour un même système.
Dans ce modèle, contrairement à d'autres modèles de dégradation tels que ARD_∞ [2], l'effet de la maintenance ne prend en compte que la dégradation qui survient depuis la dernière maintenance et non depuis le début de la dégradation.
Le modèle ARD 1 a trois paramètres à estimer : Le paramètre ρ d'efficacité de la maintenance, le paramètre μ compris dans l'espérance de la loi des accroissements et le paramètre σ^2 compris dans la variance de ces accroissements.

L'observation ou non des niveaux de dégradation au moment des maintenances amène à distinguer plusieurs cas de figure. Les fonctions de vraisemblances et les estimations des paramètres du modèle seront différents d'un cas à l'autre.

2 Inférence statistique

Le système considéré fait l'objet d'inspections, pendant lesquelles on peut mesurer le niveau de dégradation. Des mesures peuvent être effectuées aux instants des maintenances, juste avant et juste après, et en dehors des instants de maintenance.

Quatre cas se dessinent :

- Premier cas : on observe le niveau de dégradation juste avant et juste après la maintenance.

- Second cas : on n'observe pas la dégradation juste après la maintenance mais on l'observe juste avant.
- Troisième cas : on n'observe pas la dégradation juste avant la maintenance, en revanche on l'observe juste après.
- Quatrième cas : on n'observe la dégradation ni juste avant ni juste après la maintenance.

Ces quatre situations impliquent différentes écritures de la vraisemblance du modèle et a fortiori différentes estimations des paramètres.
On considérera ici que pour un même système, toutes les maintenances admettent le même cas d'observation.
Par la suite, on appellera "saut" la différence entre le niveau de dégradation juste avant et juste après une maintenance.

Dans la figure ci-dessous, pour chacun des cas les mesures et les maintenances sont effectuées périodiquement. Lorsqu'une mesure n'est pas effectuée lors d'une action de maintenance, le saut apparaît alors en pointillé.

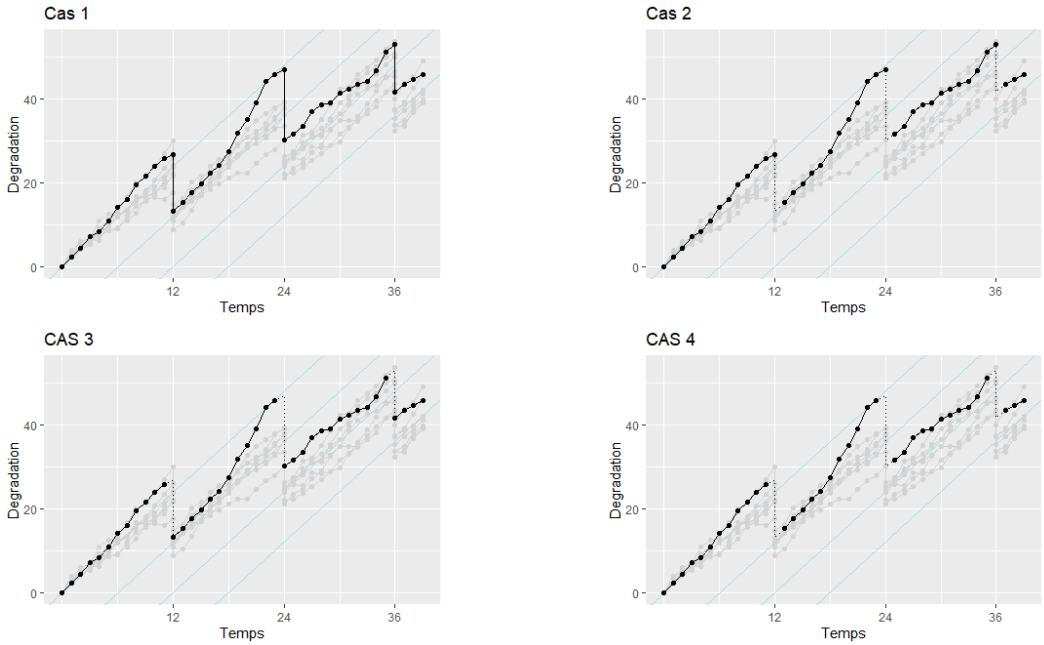


Figure 1: Processus de dégradation avec maintenances imparfaites suivant les quatre cas d'observation

Dans le premier cas, les deux niveaux de dégradation juste avant et juste après les maintenances sont observées (cf. cas 1 figure 1). Le saut de la maintenance est alors connu et déterministe. La vraisemblance dans ce cas-ci s'écrira alors comme un produit de densités de lois normales indépendantes et d'une distribution de Dirac.
Le saut étant déterministe, on estimera ici seulement les paramètres μ et σ^2 par maximum de vraisemblance.

Dans le second cas, le niveau de dégradation juste après la maintenance n'est pas observé (cf cas 2 figure 1). Cependant, le saut s'écrit en fonction des incréments précédents depuis la dernière maintenance. Donc, puisqu'on connaît la valeur initiale du processus (qui vaut 0 au temps 0) et que l'on connaît la valeur de la dégradation juste avant la première maintenance, on peut en déduire la loi conditionnelle du premier saut. Puis, par récurrence, on peut déterminer les lois conditionnelles de tous les sauts, et en déduire l'expression de la fonction de vraisemblance.

Le troisième cas est plus simple (cf cas 3 figure 1), puisque l'incrément non observé juste avant la maintenance est indépendant des autres incréments. Cela permet d'obtenir facilement les lois conditionnelles des sauts.

Le quatrième cas est plus complexe (cf cas 4 figure 1). Chaque saut dépendant des incréments précédents depuis la dernière maintenance, l'incrément non observé juste après chacune des maintenances se retrouve dans l'écriture du saut suivant. Ainsi, les lois conditionnelles des sauts sont plus complexes et font intervenir des vecteurs gaussiens. On arrive néanmoins à écrire la fonction de vraisemblance.

Dans tous les cas, on montre que les estimateurs de maximum de vraisemblance de μ et σ^2 s'expriment en fonction de celui de ρ. L'estimateur de maximum de vraisemblance de ρ est déterminé comme solution d'une équation implicite.

A l'aide de simulations, on comparera les quatre situations évoquées et on évaluera la qualité des estimateurs dans chacun des cas.

Bibliographie

[1] Mercier, S. et Castro, I.T., *Stochastic comparisons of imperfect maintenance models for a gamma deteriorating system*, European Journal of Operational Research, 273 (2019), 237–248.

[2] Salles, G., Mercier, S. et Bordes, L. *Semiparametric estimate of the efficiency of imperfect maintenance actions for a gamma deteriorating system*, Journal of Statistical Plan-

ning and Inference, 206, (2020), 278-297.

La place de la statistique dans les nouveaux programmes du baccalauréat 2021

Frédérique Letué[1] & Anne-Béatrice Dufour [2] & Antoine Rolland [3]

[1] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, Frederique.Letue@univ-grenoble-alpes.fr*
[2] *Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, 43 Bd du 11 novembre 1918, F-69622 Villeurbanne, France, anne-beatrice.dufour@univ-lyon1.fr*
[3] *Université de Lyon, Université Lyon 2, Laboratoire ERIC, Campus Porte des Alpes – 5 av. Pierre Mendès-France, 69676 Bron Cedex, France antoine.rolland@univ-lyon2.fr*

Résumé. Le baccalauréat français a subi une des plus importantes réformes de ces dernières années. Suite à la suppression des séries dans la voie générale, l'enseignement de mathématiques (et par conséquent de la statistique) joue un rôle particulier dans cette réforme par son absence dans le tronc commun de la filière générale et sa présence sous la forme d'une spécialité en Première et Terminale et de deux options (mathématiques expertes et mathématiques complémentaires) en Terminale. Dans cet exposé, après une présentation générale de la réforme et des nouveaux programmes de mathématiques, nous montrons quelle place tient la statistique, en distinguant la voie générale et la voie technologique, les spécialités et les options, son apparition éventuelle dans d'autres disciplines.

Mots-Clés. Enseignement de la statistique, Réforme du baccalauréat

Abstract. The French « baccalauréat » was submitted to one of the main reforms in the latest years. Following the cancellation of the specializations in the general field, mathematical teaching (and thus statistics teaching) plays a singular place in the reform, since it does not appear in the common-core syllabus, but it appears as a specialization teaching in the « Première » and « Terminale » years, and also as two options (« expert mathematics » and « complementary mathematics ») in the « Terminale » year. In this talk, after presenting the reform and the new mathematics syllabus, the new place of statistics teaching is shown, distinguishing between technologic and general fields, specialization teachings and options, and possible occuring in other disciplines.

Keywords. Statistics teaching, French high-school diploma reform

1. La réforme du baccalauréat 2021

Le baccalauréat a subi une des plus importantes réformes de ces dernières années. Les séries générales littéraire, économique et sociale, scientifique (L, ES et S) ont été supprimées[1]. En remplacement, les lycéens ont un tronc commun de 16h par semaine comprenant du français en Première, de la philosophie en Terminale, de l'histoire géographie, un enseignement moral et civique, deux langues vivantes, de l'éducation physique et sportive, un enseignement scientifique pour la filière générale et des mathématiques pour la filière technologique. A côté

[1] https://www.education.gouv.fr/en-route-vers-le-baccalaureat-2021-le-ministre-la-rencontre-de-parents-d-eleves-de-seconde-9584

de ce tronc commun, les lycéens ont à choisir 3 spécialités[2] de 4h par semaine en Première et 2 spécialités de 6h par semaine (parmi les 3 choisies en Première) en Terminale :

Arts	Biologie écologie (pour les lycées agricoles uniquement)
Histoire-géographie, géopolitique et sciences politiques	Humanités, littérature et philosophie
Langues, littératures et cultures étrangères et régionales	Littérature et langues et cultures de l'Antiquité
Mathématiques	Numérique et sciences informatiques
Physique-chimie	Sciences de la vie et de la Terre
Sciences économiques et sociales	Sciences de l'ingénieur

Deux options au maximum peuvent éventuellement être ajoutées en voie générale : une troisième langue vivante, arts, éducation physique et sportive, ou langues et cultures de l'Antiquité dès la Première, mathématiques expertes, mathématiques complémentaires, droit et grands enjeux du monde contemporain en Terminale.

En voie technologique, le tronc commun est le même que celui de la voie générale, à part l'enseignement scientifique qui est remplacé par des mathématiques (3h par semaine). Les spécialités dépendent des filières choisies, qui sont, elles, maintenues.

Les épreuves du baccalauréat sont elles-mêmes modifiées : si les épreuves anticipées de français en fin de classe de Première et celle de philosophie en Terminale sont maintenues, les deux spécialités de Terminale quant à elles sont évaluées par deux épreuves et via un grand oral portant sur ces deux spécialités en fin de Terminale. Ces épreuves comptent pour 60% de la note finale du baccalauréat. Les autres disciplines de tronc commun sont évaluées via des épreuves communes de contrôle continu (appelées E3C en 2019-2020, puis renommées EC en 2020-2021) en Première (deux sessions) et en Terminale (une session). La spécialité abandonnée en Terminale est évaluée en fin de Première. Les notes à ces épreuves constituent 30% de la note finale. Enfin, les notes des bulletins de Première et Terminale fournissent les 10% restants, permettant ainsi d'inclure l'évaluation des options.

La mise en place de la réforme, prévue sur les deux années scolaires 2019-2020 et 2020-2021, a été fortement perturbée, d'une part à cause de la contestation de la réforme elle-même ayant entraîné des blocages de lycées lors des premières épreuves en 2019-2020, mais surtout en raison de la crise sanitaire de la COVID-19. Ainsi, sur les trois séries d'épreuves communes prévues, seule la première a réellement eu lieu en janvier-février 2020, les deuxième et troisième étant annulées et remplacées par la note de contrôle continu. De même, les épreuves anticipées de français, l'épreuve de spécialité de fin de Première, et les deux épreuves de spécialités de Terminale (qui devaient se tenir en mars 2021) ont été supprimées. A la date où nous écrivons, sont toujours prévues les épreuves de philosophie et le grand oral.

2. Les mathématiques dans la réforme du baccalauréat 2021

Les mathématiques jouent un rôle particulier dans la réforme du baccalauréat 2021 par leur absence dans le tronc commun de la voie générale et par leur présence dans deux options

[2] https://www.education.gouv.fr/les-programmes-du-lycee-general-et-technologique-9812

(mathématiques expertes et mathématiques complémentaires). Les mathématiques de Seconde sont les mêmes pour les voies générale et technologique à raison de 4h par semaine (voir « Programmes et ressources d'accompagnement pour les voies générale et technologique du lycée » du Ministère de l'Education Nationale[3]).

2.1. Le cas du baccalauréat technologique

Les mathématiques sont présentes dans le tronc commun de la voie technologique à raison de 3h par semaine en Première comme en Terminale. Les programmes de Première et de Terminale portent essentiellement sur le vocabulaire ensembliste et logique, les automatismes, l'analyse et la statistique et probabilités. La série sciences et technologies du design et des arts appliqués (STD2A) se différencie des autres filières par l'introduction d'activités géométriques en lieu et place de l'algorithmique et programmation présente dans toutes les autres filières Sciences et Techniques.

2.2. Le cas du baccalauréat général

Dans la voie générale, les mathématiques ne sont pas présentes dans le tronc commun (hormis quelques apartés dans l'enseignement scientifique). En Première, si le lycéen a choisi la spécialité « mathématiques », il en suit 4h par semaine, sinon 0h. En Terminale, cela dépend de la spécialité conservée et des options et peut ainsi varier de 0h à 9h.
Le programme de la spécialité de Première s'appuie sur les acquis de la Seconde et porte sur cinq grands thèmes : l'algèbre, l'analyse, la géométrie, les probabilités et la statistique ainsi que sur l'algorithmique et programmation. Le programme de la spécialité de Terminale s'appuie sur les acquis de la Seconde et de la Première avec un approfondissement des connaissances et l'acquisition d'un niveau de compétences conduisant à la préparation de l'enseignement supérieur. Cinq termes sont abordés : algèbre et géométrie, analyse, probabilités, algorithmique et programmation et l'acquisition d'un vocabulaire ensembliste et logique.

3. La statistique dans les programmes de mathématiques

La statistique avait pris une place importante dans les dernières versions des programmes de mathématiques. Le but de ce paragraphe est de voir quelle place elle occupe maintenant, en distinguant la voie générale et la voie technologique, les spécialités et les options, son apparition éventuelle dans d'autres disciplines.

3.1. Le socle commun de la Seconde

Le programme de mathématiques comme de statistique de Seconde est commun aux filières technologique et générale et s'appuie sur les acquis du cycle 4 de formation. Il prévoit une consolidation autour de la notion de pourcentage vue soit comme une proportion, soit comme une évolution. La statistique descriptive qui proposait l'étude de trois paramètres (moyenne, médiane, étendue) s'enrichit de la moyenne pondérée ainsi que de deux paramètres de dispersion : écart interquartile et écart-type. En probabilités, l'accent est mis sur la formalisation de la notion de lois de probabilités dans le cas fini en s'appuyant sur le langage des ensembles

[3] https://cache.media.eduscol.education.fr/file/SP1-MEN-22-1-2019/95/7/spe631_annexe_1062957.pdf

et les premiers éléments de calcul de probabilités. Les notions d'échantillonnage et de loi des grands nombres sont présentées sous une forme expérimentale en lien avec la partie algorithmique et programmation.

3.2. Le baccalauréat technologique

En Première, l'accent est mis sur les couples de variables catégorielles, le croisement des informations. Les probabilités s'enrichissent de la notion de probabilité conditionnelle, de modèle associé aux expériences aléatoires par l'utilisation de simulations, l'aspect formalisme étant mis de côté. En Terminale, l'accent est mis sur les séries à deux variables quantitatives : nuage de points, changements de variable et le modèle de la régression linéaire simple (ajustement affine et moindres carrés). La progression se poursuit autour des probabilités conditionnelles avec les notions d'indépendance de deux événements, de la formule des probabilités totales. La notion de variable aléatoire discrète à valeurs finies est introduite à travers la loi binomiale, le calcul de l'espérance ainsi que les différents attendus liés aux coefficients binomiaux.

3.3. Le baccalauréat général

Dans le cadre du baccalauréat général, on distingue d'une part les deux enseignements de spécialité de Première et Terminale, d'autre part l'enseignement optionnel de mathématiques complémentaires de Terminale. L'autre option « mathématiques expertes » ne contient ni probabilité, ni statistique, à part l'étude des chaînes de Markov en lien avec les matrices.

L'enseignement de spécialité de Première s'attache à l'étude des notions suivantes : probabilité conditionnelle, indépendance et variables aléatoires réelles (loi, espérance, variance) dans le cadre des univers finis. Les modèles probabilistes sont approfondis en Terminale. Le schéma de Bernoulli se décline en la somme d'épreuves identiques et indépendantes et permet ainsi d'aborder les sommes de variables aléatoires et les différentes relations inhérentes à ces sommes pour les espérances et les variances. Enfin, sont abordées la loi des grands nombres et l'inégalité de Bienaymé-Tchebychev.

L'enseignement optionnel de mathématiques complémentaires proposé en Terminale traite les lois discrètes (uniforme, Bernoulli, Binomiale, géométrique), les lois à densité (uniforme, exponentielle) et la statistique à deux variables quantitatives (nuage de points, régression linéaire simple, ajustement des moindres carrés).

4. La statistique dans les autres matières

Des notions de statistique sont également présentes dans les programmes d'autres disciplines que les mathématiques. C'est le cas en particulier de la spécialité Sciences Économiques et Sociales (SES) de la filière générale, des enseignements scientifiques du tronc commun de la filière générale, et dans quelques matières en filières technologiques, telles que la physique en Sciences et Technologies de l'Industrie et du Développement Durable (STI2D) ou l'économie en Sciences et Technologies du Management et de la Gestion (STMG). Cet enseignement a pour vocation d'accroître auprès des étudiants la capacité à utiliser, à appliquer, à interpréter, à communiquer, à créer et à critiquer des informations et des idées statistiques de la vie réelle.

4.1. Spécialité « sciences économiques et sociales »

Des notions de statistique sont abordées dans le programme de la spécialité SES, tant en classe de Première qu'en Terminale, avec comme objectif affiché de servir à l'objectivité des sciences sociales. Il s'agit donc plutôt d'étude des statistiques (en lien avec la statistique publique par exemple) que d'étude de la statistique comme outil mathématique en lien avec les probabilités. Il est bien indiqué en préambule de la présentation de la spécialité : « *Les professeurs insistent sur l'exigence de neutralité axiologique. Les sciences sociales s'appuient sur des faits établis, des argumentations rigoureuses, des théories validées et non pas sur des valeurs. L'objet de l'enseignement des sciences économiques et sociales est le fruit des travaux scientifiques, transposés à l'apprentissage scolaire. Il doit aider les élèves à distinguer les démarches et savoirs scientifiques de ce qui relève de la croyance ou du dogme, et à participer ainsi au débat public de façon éclairée ; il contribue à leur formation civique ». Le tout doit être mis en œuvre « en prenant appui sur des supports variés (textes, tableaux statistiques, graphiques, utilisation de jeux, comptes rendus d'enquêtes, documents iconographiques et audiovisuels, monographies, ...)* ».

4.2. Tronc commun : enseignement scientifique

L'enseignement scientifique est une nouveauté de la réforme des lycées. Il s'adresse à tous les lycéens de Première et Terminale du baccalauréat général à raison de deux heures par semaine. Son objectif est double : une formation scientifique générale et « *un point d'appui pour ceux qui poursuivent et veulent poursuivre des études scientifiques* ». Chaque année, des thématiques relevant des sciences de la vie et de la physique-chimie sont proposées : 4 en Première suivies d'un projet expérimental et numérique et 3 en Terminale. Les mathématiques sont présentes dans chacun des thèmes ainsi que la mise en œuvre des différents concepts à travers l'outil numérique.

La statistique est présente notamment dans le thème 3 : Une histoire du Vivant. Dans le sous-thème - La Biodiversité et son évolution, les lycéens s'attachent à estimer une abondance par une méthode d'échantillonnage spécifique : Capture-Marquage-Recapture. Ils étudient également les fluctuations d'échantillonnage et l'intervalle de confiance d'une proportion. Dans le sous-thème - L'intelligence artificielle, ils travaillent autour des notions de corrélation / causalité ainsi que des modèles linéaires et exponentielles. Enfin, quelle que soit la thématique générale et dans un souci de réflexion générale sur les savoirs scientifiques, de nombreuses représentations graphiques liées à la statistique descriptive classique sont utilisées (représentations en secteurs ou en bâtons, boîte à moustaches, nuage de points, etc).

4.3. Spécialité « physique »

Le but de la spécialité physique est de familiariser le lycéen en filière STI2D avec la démarche expérimentale scientifique. A ce titre, la statistique y est présente à travers la notion de mesure d'incertitude et d'erreurs de mesure. En particulier, sont mobilisées les connaissances portant sur l'écart-type et la fluctuation d'échantillonnage. Les objectifs de compétences statistiques dans la spécialité physique sont, entre autres, d'exploiter des séries de mesures indépendantes (histogramme, moyenne et écart-type), d'estimer une incertitude-type sur une mesure unique

et de discuter de la validité d'un résultat en comparant la différence entre le résultat d'une mesure et la valeur de référence d'une part et l'incertitude-type d'autre part.

4.4. Spécialité « ingénierie, innovation et développement durable »

Dans cette spécialité (filière STI2D) résolument industrielle centrée sur la conception, un lien est fait avec la statistique à l'occasion de l'interprétation des résultats de simulations, à l'aide de « *courbe, tableau, graphe, unités associées*. ».

4.5. Spécialité « droit et économie »

La partie économie de la spécialité droit et économie de la filière STMG inclut quelques notions de statistique, toujours dans l'optique d'une lecture et d'une interprétation correcte par l'élève de tableaux statistiques. L'approche est résolument pratique : « *les élèves utilisent les notions et les mécanismes économiques à l'occasion d'analyses de situations réelles ou de données quelles qu'en soient leurs formes (séries statistiques, graphiques, cartes, etc.)* ».
Le contenu statistique en lui-même est léger et se limite à rechercher une information ou des statistiques pertinentes dans des documents fiables. Il est précisé que *« dans le cas de documents statistiques, il s'agit par exemple d'être capable d'analyser et d'interpréter des graphiques de différents formats (graphiques statistiques, hiérarchiques ou de tendances, histogrammes, nuages de points, etc.) et de mobiliser les données observées pour calculer de nouvelles statistiques (cf. valeur ajoutée, coût marginal)* ». Parmi les thèmes d'étude abordés, seuls ceux ayant trait au calcul de la richesse (calcul du PIB, notion de statistique publique, calculs d'inégalité de revenus) et au chômage (taux de chômage et taux d'emploi, au sens du BIT et de Pôle Emploi) sont orientés majoritairement autour de statistiques.

5. Discussion et conclusion

La réforme met l'accent sur la dualité de l'enseignement de la statistique. Comme le soulignent Cobb et Moore[4] (1997), la statistique est certes une discipline méthodologique mais elle existe car elle offre aux autres champs d'étude un ensemble cohérent d'idées ou d'outils pour traiter les données. Une partie de l'enseignement en lycée est donc axée sur son lien avec la mathématique (§ 3) ; l'autre partie sur l'analyse de données impliquée dans un domaine (§ 4). Il s'agit dans ce dernier cas de défendre une numératie[5] statistique auprès de tous.
Alors que les programmes de mathématiques au lycée faisaient précédemment la part belle à la statistique et à l'utilisation pratique de fichiers de données réelles, l'esprit des programmes de spécialités de mathématiques en Première et Terminale générale est maintenant beaucoup plus orienté vers les aspects probabilistes et historiques. Seule l'option mathématiques complémentaires de Terminale affirme le côté appliqué de la statistique avec de nombreux liens faits avec d'autres disciplines et laisse une belle place à la statistique à deux variables (droite de régression, coefficient de corrélation), peu présente auparavant. La notion de fluctuation d'échantillonnage est, elle, nettement moins présente dans les programmes actuels.

[4] Cobb, G.W. and Moore D.S. (1997) Mathematics, Statistics and Teaching. Am. Math. Mon, 104(9), 801:823

[5] capacité à utiliser, à appliquer, à interpréter, à communiquer, à créer et à critiquer des informations et des idées mathématiques de la vie réelle

Classification supervisée par arbre binaire et modèle linéaire généralisé

Lorena León [1,2], Jean Peyhardi [2], Catherine Trottier [2,3]

[1] *CIRAD, UMR AGAP, Montpellier, France.*
yinneth-lorena.leon_velasco@cirad.fr
[2] *IMAG, Univ Montpellier, CNRS, Montpellier, France.*
jean.peyhardi@umontpellier.fr
[3] *Univ Paul Valéry Montpellier 3, Montpellier, France.*
catherine.trottier@umontpellier.fr

Résumé. La représentation hiérarchique des données est une approche intéressante dans le cadre de la régression réponse catégorielle, et ainsi en classification supervisée. Certaines des catégories de réponse sont souvent incluses dans d'autres, ce qui génère des subdivisions successives. Cette structure est proprement représentée par un dendrogramme ; dans lequel pour chaque nœud interne, l'estimation d'un modèle (utilisant un ensemble spécifique de covariables) permet d'extraire des informations plus fines sur la différenciation des catégories. Dans la plupart des cas, cette structure est *a priori* inconnue. Dans cet article, nous introduisons et illustrons une méthodologie qui permet de trouver un arbre de partition binaire pour les J catégories réponses, suivi d'une estimation de la fonction de lien ainsi que de la sélection de covariables pertinentes pour chacune des partitions présentées dans l'arbre. Nous avons testé notre méthodologie sur différents jeux de données benchmarks et nous avons comparé d'une part la qualité de la méthode de recherche d'arbre mais aussi sa performance face au logit multinomial classique.

Mots-clés. Structure hiérarchique des catégories, arbre de partition, fonction de lien, spécification des GLMs, dendrogramme, PCGLM

Abstract. The hierarchical representation of data can be meaningful within the framework of regression for categorical responses and thus for supervised classification. Some of the response categories are likely to enclose others leading to successive subdivisions. This structure is appropriately represented by a dendrogram; where for each internal node, the estimation of a model (using a specific set of covariates) allows for the extraction of finer information on category differentiation. In most cases, this structure is *a priori* unknown. In this paper, we introduce and illustrate a methodology to find a binary partition tree for the J categories of a response variable, followed by an estimation of the link function together with the selection of pertinent covariates for each of the partitions represented in the tree. We tested our methodology on different datasets and we compared its performance against the classical multinomial logistic model.

Keywords. Hierarchical structure among categories, partition tree, link function, specification of GLMs, dendrogram, PCGLM.

1 PCGLM binaire

Considérons le cadre de régression dans lequel la réponse Y est une variable catégorielle à J catégories, qui doit être expliquée par un ensemble de p variables explicatives $\boldsymbol{x} = (x_1, \ldots, x_p)$. Dans ce contexte, un arbre de partition peut être employé pour spécifier la hiérarchie entre les catégories réponses. Pour ce travail, nous ne considérons que des arbres de partition binaires ce qui simplifie la spécification du modèle à construire. Dans la suite, nous introduisons les notations et définitions nécessaires pour le développement de cette méthodologie.

Un arbre orienté $\mathcal{T}$ est un *arbre de partition de* $\{1, \ldots, J\}$ où :

- les nœuds frères constituent une partition non identique de leur nœud père,
- $\{1, \ldots, J\}$ est la racine de $\mathcal{T}$ et
- chaque catégorie $\{j\}$ (les feuilles) fait partie de $\mathcal{T}$.

Un arbre orienté $\mathcal{T}$ est considéré un *arbre de partition binaire* si, en plus des conditions précédentes, le degré de chaque noeud est au maximum 2. Désormais, $\mathcal{V}^*$ est l'ensemble des nœuds non terminaux d'un arbre de partition binaire $\mathcal{T}$.

Il est possible de représenter un arbre de partition binaire par un *dendrogramme*. Dans notre contexte, nous ne considérons que les dendrogrammes *étiquetés, non ordonnés (E-NO)* qui sont équivalent à une arbre de partition binaire. Le nombre de dendrogrammes E-NO définis pour les J catégories est donné par la formule :

$$b(J) = \frac{(2J-2)!}{2^{J-1}(J-1)!} \tag{1}$$

À titre d'exemple, dans le cas où $J = 5$, la Figure 1 représente les trois structures de dendrogrammes possibles. Il y a 60 étiquetages possibles pour le dendrogramme (i), 30 pour le dendrogramme (ii), et 15 pour le dendrogramme (iii).

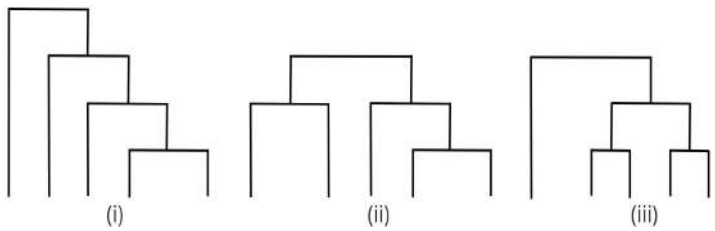


Figure 1: Structures de dendrogrammes étiquetés et non classés pour 5 catégories

Table 1: Nombre de dendrogrammes selon J.

J	3	4	5	6	7	8	9	10
b(J)	3	15	105	945	10395	135135	2027025	34459425

Pour plus de 5 catégories, le nombre de dendrogrammes augmente à tel point qu'il devient difficile d'explorer toutes les possibilités (voir Tableau 1).

Un **GLM conditionnel partitionné binaire de** $\{1, \ldots, J\}$ (B-PCGLM, pour son acronyme en anglais) est spécifié par :

- un arbre de partition binaire $\mathcal{T}$ de $\{1, \ldots, J\}$ (dendrogramme E-NO)
- une collection de modèles de régression binaire $\mathcal{C}$, où pour chaque nœud $v \in \mathcal{V}^*$, F^v est la fonction de lien associée, et $\boldsymbol{x}^{\boldsymbol{v}}$ est le sous-vecteur de l'ensemble des covariables $\boldsymbol{x}$, associé au nœud v.

Étant donné la structure hiérarchique d'un B-PCGLM, il est possible de faire des estimations séparés des $J-1$ modèles. La vraisemblance globale se décompose comme suit :

$$l = \sum_{v \in \mathcal{V}^*} l^v$$

où l^v représente la vraisemblance du GLM binaire associé au noeud v. Dans la section suivante, nous présentons une méthodologie pour construire un B-PCGLM en supposant que la structure du dendrogramme est inconnue.

2 Construction d'un B-PCGLM

Pour obtenir un B-PCGLM, deux tâches principales doivent être réalisées. La première est la définition de la structure hiérarchique des catégories de réponse. À notre connaissance, il n'existe pas de méthodologie permettant de générer automatiquement une structure hiérarchique pour les groupes établis. Ici, nous proposons une heuristique pour trouver le meilleur dendrogramme pour la réponse dans un contexte de régression catégorielle. La deuxième tâche consiste à trouver et à ajuster l'ensemble des modèles $\mathcal{C}$ qui génèreront des informations spécifiques pour chacun des $J-1$ nœuds.

2.1 Construction d'un arbre de partition binaire

L'inférence de tous les dendrogrammes possibles peut s'avérer très couteuse en temps de calcul lorsque le nombre de catégories augmente (car le nombre de dendrogrammes explose lorsque J augmente; voir tableau 1). Nous proposons alors une heuristique de construction

du dendrogramme, basée sur la classification ascendante hiérarchique (CAH) mais dans un contexte particulier. Au lieu de regrouper les individus, nous considérons comme points de départ les groupes $E_j := \{1 \leq i \leq n : y_i = j\}$ pour $j = 1, \ldots, J$, puis nous procédons à une série de fusions successives des groupes jusqu'à ce qu'ils soient tous membres d'un seul et même groupe, la racine.

Matrice de dissimilarité (entre les individus) D

Pour trouver les distances entre les individus selon les p covariables, une première étape consiste à transformer toutes les covariables quantitatives sur une échelle commune. Si elles ont le même niveau d'importance, la procédure la plus appropriée consiste à standardiser la covariable k, en la divisant par son rang $r_k = \max_{1\leq i\leq n} x_{i,k} - \min_{1\leq i\leq n} x_{i,k}$.

Pour les variables quantitatives, les distances Euclidiennes et de Manhattan sont les plus populaires, tandis que pour les variables ordinales et nominales, les dissimilarités les plus connues sont respectivement celles de Bray-Curtys et de Sokal-Michener; pour plus de détails voir (Deza et al., 2013). L'approche la plus appropriée pour aborder les variables mixtes est d'utiliser la distance de Gower, c'est-à-dire pour deux individus i et i' nous avons :

$$D_{i,i'} = \frac{1}{p}\sum_{k=1}^{p} d_{i,i'}^k$$

avec $d_{i,i'}^k = \frac{|x_{i,k}-x_{i',k}|}{r_k}$ si la $k^{\text{ième}}$ covariable est quantitative, et $d_{i,i'}^k = \mathbf{1}_{\{x_i \neq x_{i'}\}}$ (fonction indicatrice) si elle est catégorielle. La matrice D est de dimension $n \times n$.

Matrice de dissimilarité (entre les groupes) Δ

À partir de la matrice des dissimilarités entre les individus D, il est nécessaire de trouver les dissimilarités entre les J groupes $E_1, \ldots, E_J$. Il existe plusieurs définitions des dissimilarités entre les groupes et leurs méthodes associées sont connues sous le nom de méthodes de liaison, parmi les plus populaires : minimum, maximum et moyenne. Par exemple, la dissimilarité de liaison minimum est celle de la plus proche paire d'individus, où les paires sont composées d'un individu de chaque groupe :

$$\Delta_{j,j'} := \min_{i\in E_j, i'\in E_{j'}} D_{i,i'} \ .$$

La matrice Δ est de dimension $J \times J$.

Regroupement hiérarchique à partir de Δ

En partant des groupes $E_1, \ldots, E_J$, de nouveaux groupes sont créés séquentiellement comme l'union des sous-groupes les plus similaires. C'est le même principe que celui employé dans le CAH (Everitt et al.). Cela permet de créer le dendrogramme E-NO recherché qui résume la structure hiérarchique qui reflète les différences et/ou les similitudes entre les catégories de réponse.

2.2 Modèles binaires pour chaque nœud interne

Choix des fonctions de lien

Parmi les fonctions de lien les plus populaires pour la modélisation des données de réponse binaire, il y a les fonctions de distribution logistique et normale. Une alternative moins commune est la distribution Student $t(\nu)$ à ν degrés de liberté, qui a prouvé être une alternative robuste pour les modèles de régression (Lange et al., 1989). Liu (2004) a démontré que le modèle avec la fonction de lien $t(7)$, fournit une excellente approximation du lien logistique ; et qu'avec un grand nombre de degrés de liberté, le modèle se rapproche du modèle Probit. Ainsi, les distributions de Student avec $\nu > 0$ offrent un ensemble infini et remarquable de possibilités pour la fonction de lien.

Pour différentes expérimentations sur des données réelles ainsi que sur des simulations, nous avons trouvé le profil de vraisemblance pour $\nu \in (0.25, 30)$. Sur cette base, nous avons mis au point une approche heuristique pour trouver le ν qui correspond au meilleur modèle en termes de vraisemblance : estimer deux modèles en utilisant les distributions logistique et Cauchy comme fonctions de lien, si la vraisemblance du modèle logistique est supérieure ou égale à celle de Cauchy, alors estimer le modèle Probit et garder le modèle avec la vraisemblance la plus élevée ; dans l'autre cas, utiliser un algorithme d'optimisation pour trouver le meilleur $\nu \in (0.25, 4)$.

Sélection de variables pour chaque nœud interne

Comme les $J - 1$ modèles doivent être estimés à partir du regroupement de catégories, la question de la sélection des variables explicatives qui influencent chaque nœud se pose. Dans notre méthodologie, nous avons proposé d'utiliser l'approche Lasso qui effectue à la fois : la régularisation et la sélection des variables. Il convient de noter que pour s'assurer que la fonction de lien correspond bien aux données, après avoir sélectionné les covariables pour chaque nœud, il est recommandé de réestimer la fonction de lien en utilisant l'ensemble des covariables sélectionnées x^v.

3 Application

Nous avons testé la méthodologie présentée sur plusieurs jeux de données de référence. L'un d'entre eux consiste à déterminer le type de maladie *Eryhemato-Squamous* parmi les six catégories suivantes : *psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis* et *pityriasis rubra pilaris*. Pour les variables explicatives, on dispose de 12 attributs cliniques et 22 caractéristiques histopathologiques. Le problème de classification est assez difficile car les caractéristiques cliniques observées sont similaires. Afin de mesurer la qualité du B-PCGLM obtenu, nous l'avons comparé aux différents arbres binaires possibles. Avec 6 catégories réponses, il existe un total de 945 arbres différents (cf Formule 1). Nous avons estimé l'ensemble des 945 arbres en utilisant la distribution logistique comme fonction de lien (afin de comparer uniquement les strutures d'arbres).

En ordonnant les erreurs de classification de ces arbres, nous avons constaté par validation croisée que le score du B-PCGLM se situait dans des erreurs de classification les plus faibles (premier quartile). En ce qui concerne le modèle multinomial, le pourcentage moyen d'erreur de classification est de $4,49\%$, tandis que pour le B-PCGLM (après une sélection du degré de liberté de la fonction de lien Student et des variables explicatives à chaque noeud), on obtient une erreur de classification de $3,12\%$.

Nous avons aussi testé notre méthodologie sur plusieurs autres jeux de données. Dans chaque cas, nous avons observé :

- la qualité du dendrogramme parmi l'ensemble des possibilités, et
- la qualité du modèle B-PCGLM résultant en comparaison avec le modèle logit multinomial.

Pour ces deux points, les résultats sont satisfaisants, même lorsque le nombre de catégories est élevé. Nous avons donc fourni une méthodologie fiable qui permet d'éviter l'exploration de l'espace des dendrogrammes. Nous obtenons ainsi une structure hiérachique pertinente de la variable réponse, permettant un ajustement précis du modèle sur chaque nœud non-terminal.

Bibliographie

Peyhardi, J. and Trottier, C. and Guédon, Y. (2016). *Partitioned conditional generalized linear models for categorical responses*. Statistical Modelling, 16, 297 – 321.
Murtagh, F. (1984). *Counting dendrograms: A survey*. Discrete Applied Mathematics, 7, 191 – 199.
Everitt, B.S. and Landau, S. and Leese, M. and Stahl, D. et al. (2011). *Cluster Analysis, Fifth Edition (Wiley Series in Probability and Statistics)*. Wiley, 5th.
Lange, K., Roderick J. A. Little, amd Jeremy M. G. Taylor. (1989). *Robust Statistical Modeling Using the t Distribution*. Journal of the American Statistical Association, 84(408), 881 – 896.
Liu, C. (2004). *Robit regression: a simple robust alternative to logistic and probit regression*. Applied Bayesian Modeling and Casual Inference, 227 – 238.
Deza, M. and Deza, E. (2013). *Encyclopedia of Distances*. Springer Berlin Heidelberg.

Semi-Parametric Wavefront Modelling for the Point Spread Function

Tobias Liaudat [1], Jean-Luc Starck [1] & Martin Kilbinger [1]

[1] *AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, F-91191 Gif-sur-Yvette, France*
{*tobias.liaudat, jean-luc.starck, martin.kilbinger*}*@cea.fr*

Résumé. Nous présentons une nouvelle approche pour estimer le champ de la fonction d'étalement du point d'un télescope optique en construisant un modèle semi-paramétrique de son erreur de front d'onde. Cette méthode est particulièrement avantageuse car elle ne nécessite pas d'observations de calibration pour récupérer l'erreur de front d'onde et elle prend naturellement en compte la chromaticité du système optique. Le modèle est différentiable de bout en bout et s'appuie sur un opérateur de diffraction qui nous permet de calculer les fonction d'étalement du point monochromatiques à partir des informations du front d'onde.

Mots-clés. Modélisation de la Fonction d'Étalement du Point, Traitement d'Images, Optique, Lentille gravitationnelle Faible.

Abstract. We introduce a new approach to estimate the point spread function (PSF) field of an optical telescope by building a semi-parametric model of its wavefront error. This method is particularly advantageous because it does not require calibration observations to recover the wavefront error and it naturally takes into account the chromaticity of the optical system. The model is end-to-end differentiable and relies on a diffraction operator that allows us to compute monochromatic PSFs from the wavefront information.

Keywords. Point Spread Function Modelling, Image Processing, Optics, Weak Lensing.

1 Introduction

Future cosmological surveys will require to measure the shape of galaxies with high accuracy. However, the optical instruments used inevitably affect the observations with the point spread function (PSF). If we do not correct the images for the PSF, we will have significantly biased shape measurements resulting in biased cosmological analyses. Thus, the crucial importance of building a reliable and precise PSF model that allows us to take into account the PSF effects in the shape measurement. Next-generation imaging surveys, like the Euclid space mission [14], are pushing the limits of weak gravitational lensing experiments [10] providing the motivation of this study.

There exist two main approaches to PSF modelling, parametric and non-parametric. The first one consists in building an optical model of the telescope. The model is described

by a reduced number of parameters that express a high variability. This approach requires a good knowledge of the optical system physics and normally is based on the reconstruction of the wavefront error (WFE). The recovery of the WFE from in-focus images is a very degenerate problem with respect to the reduced number of parameters used. This explains the need of calibration information like observations of out-of-focus stars that help to break this degeneracy. The model's parameters can be modified slightly to fit the observation of in-focus stars, but the core of the model is defined previous to the observations. These models are usually reserved for space missions as the randomness added by the atmosphere would make an approach of this type unpractical. The most known example is the Tiny Tim algorithm for the Hubble space telescope [13]. The second approach relies in imaging-data to build the model with a minimal use of *a priori* information. The observed in-focus stars are considered to be samples of PSF field in the field-of-view (FoV) and are used to constrain the model. These techniques are generally based on learning features of the PSF field with a dimensionality reduction method. Then followed by an interpolation mtehod to recover the PSF at galaxy positions [9, 5, 3, 15, 8]. While the parametric models are capable of generating chromatic PSF models with complex shapes they are prone to have considerable errors if there is a mismatch between the model and the observations [7]. The non-parametric models do not suffer from the same issue as they are build on the observations. However, they experience difficulties to model the PSF chromaticity and complex PSF variations in the FoV.

In this work we present a new family of PSF modelling methods that intend to bridge the gap between the two classical approaches. We propose to build a semi-parametric model of the WFE by using an end-to-end differentiable diffraction operator. Our model is able to account for the PSF chromaticity as well as complex variations int he FoV without relying on the *a priori* information of the optical system. The non-parametric part of the WFE is able to correct for the mismatches of the parametric part and help to regularise the inverse problem of recovering the WFE from in-focus observations. We take advantage of the automatic differentiation provided by frameworks like Tensorflow [1] to build our model. As our model is completely differentiable, it is well adapted for gradient-based optimisation techniques. We expect that our model can serve as building block for novel PSF modelling methods that can profit from the power of deep neural networks (DNN) in a more physics-motivated environment.

2 Wavefront PSF modelling

Let us define the PSF field for a specific image exposure as a function that inputs a position in the instrument's focal plane (FP) coordinates and a specific wavelength, and outputs a monochromatic PSF. We denote this function $\mathcal{H} : F_P \times \mathbb{R}_+ \rightarrow I$, where $F_P \in \mathbb{R}^2$ is a FP position, and $I \in \mathbb{R}^{N \times N}$ is a square PSF postage stamp. PSF modelling consists in building an estimator, $\hat{\mathcal{H}}$, of the PSF field from observations at a set of FP positions

$\{u_i\}_{i=1,\dots,m_{obs}}$. Then, use the model to output monochromatic PSFs at another set of target FP positions $\{u_j\}_{j=1,\dots,m_{target}}$. The image $\mathcal{H}^{\lambda}_{u_j}$ corresponds to the PSF at location u_j and wavelength λ. We can model a polychromatic object observation as

$$\bar{G}_{u_j} = \int B(\lambda) \left(G^{\lambda}_{u_j} * \mathcal{H}^{\lambda}_{u_j}\right) \mathrm{d}\lambda \,, \tag{1}$$

where $\bar{G}_{u_j}$ is the polychromatic observation of the object $G^{\lambda}_{u_j}$ convolved with the PSF $\mathcal{H}^{\lambda}_{u_j}$, and $B(\lambda) \in \mathbb{R}$ is an indicator function of the instrument's passband. To model the PSF we consider certain star observations as point sources, hence samples of the PSF field. Therefore, we use them to constrain the PSF model $\hat{\mathcal{H}}$. The star observations can be approximated as

$$\bar{\mathcal{H}}_{u_i} \approx \sum_{b=1}^{N_\lambda} S_{u_i}(\lambda_b)\, \mathcal{H}^{\lambda_b}_{u_i} \,, \tag{2}$$

where we discretize the integral in (1) with N_λ wavelength bins, $S_{u_i}(\lambda_b) \in \mathbb{R}_+$ is the normalised Spectral Energy Distribution (SED) of the star at position u_i and wavelength λ_b, and we approximate $B(\lambda)$ as being an ideal band-pass filter in the instrument's wavelength range. In practice, the problem's inputs are a set of positions $\{u_i\}_{i=1,\dots,m_{obs}}$ with their SED and the corresponding observed PSF postage stamps. We have to use these inputs to constrain our PSF model that we define as follows

$$\hat{\mathcal{H}}^{\lambda}_{u_i} \propto \mathcal{D}_\theta \left\{ P \odot \exp\left[2\pi i \left(\frac{\Phi_{u_i}}{\lambda} + C^{\lambda}_{u_i} \right) \right] \right\}, \tag{3}$$

where $\odot$ is the Hadamard or element-wise product, $\mathcal{D}_\theta : \mathbb{R}^{p\times p} \to \mathbb{R}^{N\times N}$ denotes the diffraction operator with telescope-specific parameters θ, and $P \in [0,1]^{p\times p}$ and $\hat{\mathcal{H}}^{\lambda}_{u_i} \in \mathbb{R}^{N\times N}$ are the pupil function and the PSF postage stamp at wavelength λ and FP position u_i, respectively. The pupil function represents the obscurations encountered in the pupil plane, where our wavefront model is defined. The model definition is proportional as we need to rescale it so that its flux, the sum of the pixels in the postage stamp, is one. We base the diffraction operator on Fraunhofer's approximation [6] that allow us to have a rapid calculation based on the Fast Fourier Transform (FFT) algorithm. The operator is proportional to the squared absolute value of the FFT of the WFE. We take care of the variable sampling in the WFE as a function of wavelength by using a variable zero-padding of the estimated WFE before applying the FFT.

The parametric model, Φ_{u_i}, is based on a weighted sum of Zernike polynomials [17]. An interesting property of these polynomials is that they are orthogonal in the unit disk making them well adapted to describe mirror aberrations. Our model reads

$$\Phi_{u_i}[x,y] = \sum_{j=1}^{N_Z} a_{j,u_i} Z_j[x,y] \,, \tag{4}$$

where $a_{j,u_i} \in \mathbb{R}$ corresponds to the coefficient for the Zernike polynomial of Noll's single-index j [17], $Z_j \in \mathbb{R}^{p\times p}$ to the disk of the Zernike polynomial of index j, and N_Z is the maximum index considered. The $[x, y]$ are coordinates of the pupil plane and are not to be confused with the FP coordinates u_i. To model the variations on the FP we define a_{j,u_i} as a polynomial of the FP positions. Each coefficient with Zernike index j has an independent position polynomial of maximum order d_P. For example, a maximum degree of 1 gives us $a_{j,u_i} = c_0^j + c_1^j u_i[0] + c_2^j u_i[1]$, where $u_i[0]$ and $u_i[1]$ are the first and the second coordinates of the FP position, respectively. The number of parameters of Φ to estimate is $N_Z(d_P+1)(d_P+2)/2$. The chromatic variation of the parametric model is encoded in the λ divisor of Φ_{u_i} in (3) which corresponds to the natural chromatic variations due to diffraction.

The non-parametric model, $C_{u_i}^{\lambda}$, is defined with a simple, but successful, model in this work. We use a matrix factorisation scheme where each FP position is a weighted sum of learned WFE features. The weights are constructed as polynomials of the FP position with a maximum degree d_{NP} providing r coefficients. The model reads

$$C_{u_i}^{\lambda} = \frac{1}{\lambda}(\Pi_{u_i} \times_1 S)_{i_2,i_3} = \frac{1}{\lambda}\sum_{i_1=1}^{r}(\Pi_{u_i})_{i_1} S_{i_1,i_2,i_3} \ , \tag{5}$$

where $C_{u_i}^{\lambda} \in \mathbb{R}^{p\times p}$, $\times_1$ is the 1-mode product of a tensor with a vector [11], $S \in \mathbb{R}^{r\times p\times p}$ is the tensor containing r wavefront feature images, and $\Pi_{u_i} \in \mathbb{R}^{1\times r}$ contains the position polynomials. For example, if we set d_{NP} to 1 we obtain $\Pi_{u_i} = [1, u_i[0], u_i[1]]$. For simplicity, we only consider chromatic variations due to diffraction but the model allows to include more complex models.

3 Numerical experiment

3.1 Data

We simulate a polychromatic PSF field using the proposed model seen in (3) with its non-parametric part set to zero. The optical parameters are taken from Euclid's visible instrument characteristics [14]. We draw 200 uniformly distributed positions in the square FoV, where 70% are used for training the model and the rest for testing it. Each star is randomly assigned one of 13 stellar SED from the Pickles library [18] as done in [19, §5.3]. The maximum order of Zernike polynomial, N_Z, is set to 45, the number of wavelength bins, N_λ, to 20, the polynomial maximum degree, d_P, to 2, and the dimensions p and N to 256 and 32, respectively. The instrument's passband is considered as ideal in the range of 550 nm to 900 nm. We draw random values for the coefficients of the Zernike polynomials with the constraint that of having a RMS value of 0.1 μm to limit the optical system's aberrations. This allows to have a randomly varying PSF field. Finally, we add a random white Gaussian noise to each training star so that each one has a random SNR value

in the range $[10, 70]$. We use an Euclid-like obscuration that is composed of a circular centred obscuration with three supporting arms that can be seen in the study [20].

3.2 Experience

The PSF field modelling problem consists in estimating the simulated test stars having as input the noisy training stars. The SED information of both is available. We train and compare three models: i) Parametric model with $N_Z = 15$ and $d_P = 2$; ii) Parametric model with $N_Z = 45$ and $d_P = 2$; iii) Semi-parametric model with $N_Z = 15$, $d_P = 2$, and $d_{NP} = 3$. The objective function used to train the model is the Mean Square Error (MSE) between the observed stars and the PSF model reconstruction. All the models use a gradient-based optimization method based on the Rectified Adam optimiser [16] with a batch size of 16. Both parametric models use a learning rate of 10^{-2} and a number of epochs of 30. For the semi-parametric model, we first train the parametric part using a learning rate of 10^{-2} for 15 epochs with the non-parametric part set to zero. Then we fix the parametric part and start to train the non-parametric part using a learning rate of 1^{-1} during 100 epochs. The evaluation metric we use is the root mean square error (RMSE) between the test stars and the model estimations. Let us point out that our implementation follows the Tensorflow 2.1 framework [1] but it is used as an automatic differentiation library to perform optimisation.

3.3 Results

Table 1 presents the quantitative results. We see that the parametric model with $N_Z = 15$ is the worst performing, as expected, due to a misspecification of the parametric model with respect to the underlying data model. The parametric model with $N_Z = 45$ has a perfect match with the data model, and we observe a 35% performance gain with respect to the previous one. However, the interesting part is that the semi-parametric model with a wrongly-specified parametric part is able to obtain the best performance. Thanks to its non-parametric part the proposed model can recover the performance gap between the two parametric models, and even go beyond by obtaining a gain of almost 65% with respect to the parametric $N_Z = 45$ model.

The estimation of the Zernike coefficients using in-focus stars is a difficult problem, known to be ill-posed, and possibly presents multi-modalities. Even if we use all the training stars in our data-set we are prone to get stuck at a local minimum due to the non-convexity of the problem. The non-parametric part of the model acts as a regularizer and helps to escape from local minima by providing an easier optimization landscape. This behaviour has been observed in the optimization of overparametrized DNN and is currently a subject of study [2]. The metrics are calculated on the test stars that are not available for the training of the model, showing that the non-parametric part is not over-fitting the training stars.

Model	RMSE	Relative RMSE
Parametric $N_Z = 15$	8.69×10^{-4} (7.90×10^{-4})	11.4% (9.9%)
Parametric $N_Z = 45$	5.59×10^{-4}(4.95×10^{-4})	7.3% (6.2%)
Semi-parametric $N_Z = 15$, $d_{NP} = 3$	2.03×10^{-4} (1.71×10^{-4})	2.6% (2.1%)

Table 1: Pixel RMSE results on the test dataset, in parenthesis for the train dataset, for the different PSF models.

4 Discussion and conclusions

We have presented the building blocks that open the way to a brand-new family of PSF models that aims to improve the two classical families, which are the parametric and the non-parametric methods. The proposed semi-parametric approach allows us to reconstruct the wavefront error and derive the chromatic point spread function at any position in the field-of-view. One of the novelties of our approach is that it uses only in-focus stars to characterise the wavefront error. This is in contrast with other wavefront-based models [4, 12] that need out-of-focus stars, sometimes referred as donuts, to estimate its parameters. Nonetheless, our model could also profit from these rich observations if they are available.

The proposed PSF model provides a way to estimate the PSF field taking into account the PSF chromaticity and field-of-view variations. The semi-parametric approach is able to learn features that a parametric model cannot capture. The non-parametric part of the model corrects for a mismatch between the parametric model and the underlying observed data model. It also helps to regularize the challenging ill-posed inverse problem of estimating the wavefront error from in-focus stars. We show on a realistic data-set that it can even improve the results over a perfectly specified parametric model and that the model is not over-fitting the testing data. Our model is based on an end-to-end differentiable approach which allows for gradient-based optimization techniques which, in practice, give good results with the semi-parametric approach. Its framework allows us to easily incorporate known physical models, while the non-parametric part learns to correct for its imperfections. In this way, we can continuously incorporate physical knowledge to the model.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Tal-

war, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 6158–6169. Curran Associates, Inc., 2019.

[3] E. Bertin. Automated Morphometry with SExtractor and PSFEx. In I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots, editors, *Astronomical Data Analysis Software and Systems XX*, volume 442 of *Astronomical Society of the Pacific Conference Series*, page 435, July 2011.

[4] C. P. Davis, J. Rodriguez, and A. Roodman. Wavefront-based PSF estimation. In H. J. Hall, R. Gilmozzi, and H. K. Marshall, editors, *Ground-based and Airborne Telescopes VI*, volume 9906, pages 2156 – 2168. International Society for Optics and Photonics, SPIE, 2016.

[5] M. Gentile, F. Courbin, and G. Meylan. Interpolating point spread function anisotropy. *A&A*, 549:A1, 2013.

[6] J. W. Goodman. Introduction to fourier optics. *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, 1, 2005.

[7] S. L. Hoffmann and J. Anderson. A Study of PSF Models for ACS/WFC. Instrument Science Report ACS 2017-8, Oct. 2017.

[8] M. Jarvis et al. Dark Energy Survey Year 3 Results: Point-Spread Function Modeling. 11 2020.

[9] M. J. Jee, J. P. Blakeslee, M. Sirianni, A. R. Martel, R. L. White, and H. C. Ford. Principal component analysis of the time- and position-dependent point-spread function of the advanced camera for surveys. *Publications of the Astronomical Society of the Pacific*, 119(862):1403–1419, dec 2007.

[10] M. Kilbinger. Cosmology with cosmic shear observations: a review. *Reports on Progress in Physics*, 78(8):086901, 2015.

[11] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[12] J. E. Krist and C. J. Burrows. Phase-retrieval analysis of pre- and post-repair hubble space telescope images. *Appl. Opt.*, 34(22):4951–4964, Aug 1995.

[13] J. E. Krist, R. N. Hook, and F. Stoehr. 20 years of Hubble Space Telescope optical modeling using Tiny Tim. In M. A. Kahan, editor, *Optical Modeling and Performance Predictions V*, volume 8127, pages 166 – 181. International Society for Optics and Photonics, SPIE, 2011.

[14] R. Laureijs, J. Amiaux, S. Arduini, J.-L. Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, et al. Euclid definition study report. *ArXiv e-prints*, 2011.

[15] Liaudat, T., Bonnin, J., Starck, J.-L., Schmitz, M. A., Guinot, A., Kilbinger, M., and Gwyn, S. D. J. Multi-ccd modelling of the point spread function. *A&A*, 646:A27, 2021.

[16] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond, 2020.

[17] R. J. Noll. Zernike polynomials and atmospheric turbulence∗. *J. Opt. Soc. Am.*, 66(3):207–211, 1976.

[18] A. J. Pickles. A stellar spectral flux library: 1150 - 25000 a. *Publications of the Astronomical Society of the Pacific*, 110(749):863–878, jul 1998.

[19] M. A. Schmitz. *Euclid weak lensing : PSF field estimation.* Theses, Université Paris Saclay (COmUE), Oct. 2019.

[20] L. M. G. Venancio, L. Carminati, J. Amiaux, L. Bonino, G. Racca, R. Vavrek, R. Laureijs, A. Short, T. Boenke, and P. Strada. Status of the performance of the Euclid spacecraft. In M. Lystrup, M. D. Perrin, N. Batalha, N. Siegler, and E. C. Tong, editors, *Space Telescopes and Instrumentation 2020: Optical, Infrared, and Millimeter Wave*, volume 11443, pages 45 – 60. International Society for Optics and Photonics, SPIE, 2020.

Ψ-FPOP: un algorithme exact et rapide de segmentation avec une pénalité multi-échelle

Arnaud Liehrmann [1] [2] arnaud.liehrmann@universite-paris-saclay.fr
Guillem Rigaill [1] [2] guillem.rigaill@inrae.fr

[1] *Laboratoire de Mathématiques et Modélisation d'Évry (LAMME), Université Paris-Saclay, Université Evry, CNRS, 91037, Évry, France*
[2] *Institut des Sciences des Plantes de Paris-Saclay (IPS2), Université Paris-Saclay, Université Évry, CNRS, INRAE, 91405, Orsay, France*

Résumé. En bioinformatique, finance, traitement de la parole ou encore analyse du climat il est récurrent de traiter des séries de données dont la moyenne est sujette à une ou plusieurs ruptures. Ainsi, l'analyse de ces données implique souvent de résoudre le double problème de la détection et de la localisation de ces ruptures. Récemment, Verzelen et al. (2020) ont montré qu'il est possible de contrôler de manière optimale l'estimation de ces ruptures au sens de la perte $l1$ et de Hausdorff. La procédure implique d'optimiser un critère des moindres carrés pénalisés avec une nouvelle pénalité multi-échelle qui favorise les segmentations homogènes. Nous étendons les idées d'élagage fonctionnel à cette pénalité afin de résoudre le critère précédent. Nous proposons une implémentation efficace en C++ interfacée avec R de cet l'algorithme, baptisé Ψ-FPOP. Cette implémentation permet de traiter des grands profils rapidement.

Mots-clés. segmentation, pénalité multi-échelle, inférence par maximum de vraisemblance, optimisation discrète, programmation dynamique, élagage fonctionnel

Abstract. In bioinformatics, finance, speech processing or climate analysis it is common to process data series whose mean is subject to one or more changes. Thus, the analysis of these data often involves solving the dual problem of detecting and locating these change-points. Recently, Verzelen et al. (2020) showed that it is possible to optimally control the estimation of these change-points in the sense of $l1$ and Hausdorff loss. The procedure involves optimizing a least squares criterion penalized with a new multiscale penalty that favours well spread change-points. In order to solve the previous criterion, we extend the ideas of functional pruning to this penalty. We propose an efficient C++ implementation interfaced with R of this algorithm, named Ψ-FPOP. This implementation allows to process large profiles quickly.

Keywords. change-point detection, multiscale penalty, maximum likelihood inference, discrete optimization, dynamic programming, functional pruning

1 Définition du modèle statistique de détection de ruptures multiples

On note $Y = (y_1, \ldots, y_n)$ une série de n observations ordonnées selon un attribut. On suppose que ces observations sont des réalisations de variables indépendantes suivant une loi normale de variance constante σ^2. On assume que la moyenne de ces lois est affectée par K ruptures, ce qui correspond à une segmentation du signal en $K+1$ segments. On note τ_j la localisation de la $j^{\text{ème}}$ rupture pour $j = 1, \ldots, K$. Par convention on introduit les indices factices $\tau_0 = 0$ et $\tau_{K+1} = n$. Le $j^{\text{ème}}$ segment est formé par les observations $y_{\tau_{j-1}+1}, \ldots, y_{\tau_j}$. La moyenne de ces observations notée μ_j est un paramètre spécifique au $j^{\text{ème}}$ segment. Formellement on suppose donc le modèle suivant (Picard et al., 2007):

$$\forall i \quad | \quad \tau_{j-1} + 1 \leq i \leq \tau_j, \quad Y_i \sim \mathcal{N}(\mu_j, \sigma^2) \quad iid. \tag{1}$$

Le modèle (1) dépend d'un vecteur de paramètres $\theta = (\mu_1, \ldots, \mu_{K+1}, \sigma^2, \tau_1, \ldots, \tau_K)$. On note $\mathcal{L}(y_1, \ldots, y_n; \theta)$ la fonction de log-vraisemblance dérivée de ce modèle,

$$\mathcal{L}(y_1, \ldots, y_n; \theta) = \sum_{j=1}^{K+1} \log(f(y_{\tau_{j-1}+1}, \ldots, y_{\tau_j}; \mu_j, \sigma^2)) \tag{2}$$

avec $f(y_{\tau_{j-1}+1}, \ldots, y_{\tau_j}; \mu_j, \sigma^2)$ la distribution conjointe des observations indépendantes. Sachant l'hypothèse sur la normalité et l'indépendance des données on obtient:

$$\mathcal{L}(y_1, \ldots, y_n; \theta) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{K+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} (y_i - \mu_j)^2. \tag{3}$$

Pour $K > 0$, on définit $\mathcal{M}_{1:n}^K$ l'ensemble de toutes les segmentations possibles de Y avec exactement K ruptures ainsi que $\boldsymbol{\tau}_n = \{\tau_1, \ldots, \tau_K\} \in \bigcup_{0<K<n} \mathcal{M}_{1:n}^K$, une segmentation particulière de Y. On dénombre donc $\sum_{K=1}^{n-1} |\mathcal{M}_{1:n}^K| = \sum_{K=1}^{n-1} \binom{n-1}{K} = 2^{n-1}$ façons différentes de segmenter Y. La problématique statistique considérée ici est l'inférence du nombre et de la position des ruptures à partir des données observées. Une méthode classique est de sélectionner la segmentation $\boldsymbol{\tau}_n^* \in \mathcal{M}_{1:n}^K$ optimisant un critère de vraisemblance pénalisée.

2 Optimiser les pertes des moindres carrés avec une pénalité multi-échelle

2.1 Définition du problème d'optimisation pénalisée

Verzelen et al. ont récemment proposé une procédure contrôlant de manière optimale l'estimation des ruptures au sens de la perte $l1$ et de Hausdorff (cf. équations 31 &

32 Verzelen et al., 2020). Cette procédure repose sur l'optimisation d'un critère des moindres carrés pénalisés avec une pénalité multi-échelle. Intuitivement, la pénalité multi-échelle proposée pénalise plus fortement les petits segments et favorise les segmentations homogènes des données. Nous pouvons ré-écrire le critère à optimiser de la manière suivante:

$$\boldsymbol{\tau}_n^* = \underset{\boldsymbol{\tau}_n \in \bigcup_{0<K<n} \mathcal{M}_{1:n}^K}{\operatorname{argmin}} \left[\sum_{j=1}^{K+1} \min_{\mu} \left[\sum_{i=\tau_{j-1}+1}^{\tau j} (y_i - \mu)^2 \underbrace{-\beta \log(\tau_j - \tau_{j-1})}_{\substack{\text{pénalité dépendante de la} \\ \text{longueur du } j^{\text{ème}} \text{ segment}}} \right] + \alpha |\boldsymbol{\tau}_n| \right] \quad (4)$$

où α est de la forme $\gamma + \beta \log(n)$. γ et β sont deux constantes à calibrer.

2.2 Fonctionnalisation du problème d'optimisation pénalisée

Comme le suggère (4), le coût d'une segmentation peut-être décomposé comme la somme du coût de ses segments. On peut donc résoudre (4) à l'aide de l'algorithme de programmation dynamique de Bellman (1961) avec au maximum $\mathcal{O}(n^2)$ opérations. Si le nombre de ruptures est linéaire en le nombre de données on peut aussi utiliser l'algorithme PELT (Killick et al., 2012) qui, dans ce scénario, a une complexité typiquement linéaire. Cependant, si le nombre de rupture est petit, PELT à une complexité quadratique. Dans la suite de cette section, nous étendons les idées d'élagage fonctionnel de l'algorithme FPOP (Maidstone et al., 2017) et pDPA (Rigaill, 2015) à cette pénalité. Nous baptisons cette algorithme Ψ-Fpop. Ψ-Fpop a un temps de calcul sous-quadratique même pour un petit nombre de ruptures.

Comme dans l'article de Maidstone et al. (2017), on introduit une fonction $\widetilde{f}_{t,s}(\mu)$ qui est définie comme le coût de la meilleure segmentation des données, noté F, jusqu'au point t conditionnellement à la dernière rupture s et la moyenne du dernier segment μ. Ainsi,

$$\widetilde{f}_{t,s}(\mu) = \underbrace{F_s + \sum_{i=s+1}^{t} (y_i - \mu)^2 + \alpha}_{\text{coût fonctionnel comme dans FPOP}} - \beta \log(t-s) \quad (5)$$

avec $\widetilde{f}_{t,t}(\mu) = F_t + \alpha$ et $F_0 = -\alpha$. Durant la procédure, comme pour FPOP, cette fonction de coût est conservée et mise à jour avec la formule suivante:

$$\widetilde{f}_{t,s}(\mu) = \widetilde{f}_{t-1,s}(\mu) + (y_t - \mu)^2 + \beta \log(t-1-s) - \beta \log(t-s). \quad (6)$$

De manière analogue à FPOP, À chaque étape t, l'algorithme Ψ-FPOP considère le minimum des $\widetilde{f}_{t,s}(\mu)$, noté $\widetilde{F}_t(\mu)$, une fonction quadratique par morceaux:

$$\widetilde{F}_t(\mu) = \min_{s \leq t} \left\{ \widetilde{f}_{t,s}(\mu) \right\}. \quad (7)$$

Parmi toutes les ruptures candidates, il existe un sous-ensemble de ces ruptures qui participe à ce minimum. Notez que F_t, la solution du problème (4), s'obtient en minimisant (7) sur μ. Maidstone et al. (2017) ont démontré que $\widetilde{F}_t(\mu)$ peut être mise à jour itérativement. Dans FPOP la règle de mise à jour de jour $\widetilde{F}_t(\mu)$ suggère qu'il suffit de comparer l'ensemble des fonctions de coût des ruptures candidates participant à $\widetilde{F}_{t-1}(\mu)$ avec la fonction de coût de la rupture candidate dernièrement introduite, à savoir: $F_t + \alpha$. Ainsi, elle justifie qu'à l'étape t toutes les ruptures candidates s dont la fonction de coût $\widetilde{f}_{t,s}(\mu)$ ne participe pas à $\widetilde{F}_t(\mu)$ peuvent être élaguées. Plus formellement, pour chaque rupture candidate s on définit $Z^*_{t,s}$ l'ensemble des μ sur lesquels $\widetilde{f}_{t,s}(\mu)$ est égale à $\widetilde{F}_t(\mu)$. Sachant la règle de mise à jour de $\widetilde{F}_t(\mu)$, on obtient:

$$Z^*_{t,s} \supset Z^*_{t+1,s}. \quad \text{Et} \quad Z^*_{t,s} = \emptyset \implies Z^*_{t+1,s} = \emptyset \ . \tag{8}$$

La règle (8) n'est plus vraie pour Ψ-FPOP car $\widetilde{f}_{t,s}(\mu) - \widetilde{f}_{t,s'}(\mu)$ dépend de t. Cela implique de ré-évaluer les comparaisons entre les ruptures candidates s et s' à divers t. En particulier, on ne peut pas garantir que si $Z^*_{t,s} = \emptyset$ alors $Z^*_{t+1,s} = \emptyset$.

2.3 Mise à jour de la zone de vie des ruptures candidates ($Z_{t,s}$)

Plutôt que d'évaluer la zone de vie exacte $Z^*_{t,s}$ des ruptures candidates, nous cherchons à mettre à jour une zone de vie $Z_{t,s}$ incluant $Z^*_{t,s}$ et validant (8). Pour cela, les règles de mise à jour de $Z_{t,s}$ vers $Z_{t+1,s}$ doivent garantir que,

$$Z_{t+1,s} \supset Z^*_{t+1,s} \ . \tag{9}$$

Nous définissons l'intervalle $I_{t,s,s'}$ tel que $s < s^{'}$,

$$I_{t,s,s'} = \{\mu \mid \widetilde{f}_{t,s}(\mu) \ \leq \ \widetilde{f}_{t,s'}(\mu)\}. \tag{10}$$

La comparaison de $\widetilde{f}_{t,s}(\mu)$ avec $\widetilde{f}_{t,s'}(\mu)$ donne:

$$\widetilde{f}_{t,s}(\mu) - \widetilde{f}_{t,s'}(\mu) = \underbrace{F_s - F_{s'} + \sum_{i=s+1}^{s'} (y_i - \mu)^2}_{\textit{forme quadratique constante}} + \beta \underbrace{(\log(t - s^{'}) - \log(t - s))}_{\substack{\textit{h, une fonction qui varie} \\ \textit{avec t, s et s'}}}. \tag{11}$$

Le calcul des racines de (11) permet de déterminer $I_{t,s,s'}$. h est une fonction monotone croissante sur t et $\lim_{t\to\infty} h(t, s, s') = 0$. Comme h est croissante sur t,

$$I_{t+1,s,s'} \ \subset I_{t,s,s'} \ . \tag{12}$$

On définit alors $I_{\infty,s,s'}$ qui correspond à $I_{t,s,s'}$ lorsque $t \to \infty$. Le calcul des racines de $F_s - F_{s'} + \sum_{i=s+1}^{s'} (y_i - \mu)^2$ permet de déterminer $I_{\infty,s,s'}$.

Nous proposons la règle de mise à jour:

$$Z_{t+1,s} = Z_{t,s} \cap \overbrace{\left(\bigcap_{s'} I_{t+1,s,s'}\right)}^{\textit{comparaisons avec le futur}} \setminus \overbrace{\left(\bigcup_{s''} I_{\infty,s'',s}\right)}^{\textit{comparaisons avec le passé}} . \quad (13)$$

Nous pouvons montrer que la règle (13) valide (9) et permet d'élaguer. Intuitivement, si $Z_{t,s}$ est vide alors on peut élaguer s car pour tout $k \geq 0$ la zone de vie exacte $Z^*_{t+k,s}$ est inclue dans $Z_{t,s}$ et par conséquent est vide.

2.4 Comparaison des ruptures candidates et simplification de la règle de mise à jour

La Règle (13) suggère que pour chaque rupture candidate s, il faut la comparer à des ruptures candidates futures s' et des ruptures candidates passées s''. Pour les ruptures candidates passées il faut considérer t qui tend vers l'infini ($I_{\infty,s'',s}$). Dans ce cas, l'ensemble des comparaisons peuvent être effectuées une fois pour toute quand la rupture candidate est introduite. Pour les ruptures candidates futures il faut régulièrement comparer s à des s'. Effectuer toutes les comparaisons entre s et s' à chaque étape est coûteux: $\mathcal{O}$(nombre de ruptures candidates^2). Idéalement, pour chaque s, on aimerait effectuer le minimum de comparaisons qui entraîneraient son élagage. Une stratégie naïve (mais souvent efficace) consiste à échantilloner un petit nombre de s' à chaque étape. Notez que l'échantillonnage des s' ne change pas l'exactitude de l'algorithme.

Une implémentation de Ψ-FPOP en C++ interfacée avec R est disponible à l'adresse `https://github.com/aLiehrmann/FpopPSD`.

3 Simulations

Dans le reste de l'exposé nous présenterons des simulations portant sur le calibrage de la constante γ (cf. section 2.1), la comparaison des stratégies d'échantillonnage des ruptures candidates futures (cf. section 2.4), la comparaison des temps d'execution de Ψ-FPOP & FPOP & PELT (cf. figure 1), ainsi que la comparaison de FPOP & Ψ-FPOP sur des données réalistes de variation du nombre de copies d'ADN (CNV).

Bibliographie

Bellman, R.E., On the approximation of curves by line segments using dynamic programming, *Communications of the ACM*, 1961.
Johnson, N.A., A dynamic programming algorithm for the fused lasso and L0-segmentation, *Journal of Computational and Graphical Statistics*, 22, 246–60, 2010.

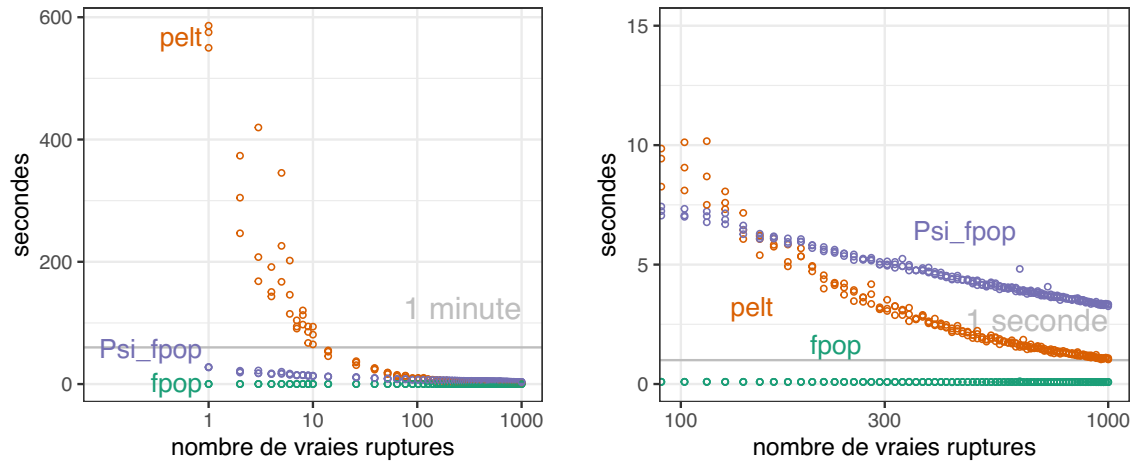


Figure 1: Temps d'exécution de PELT (pour une pénalité linéaire avec $|\boldsymbol{\tau}_n|$) & Ψ-FPOP (pour une pénalité multi-échelle) & FPOP (pour une pénalité linéaire avec $|\boldsymbol{\tau}_n|$) sur des jeux de données simulées de taille $n = 5 \times 10^5$ avec un nombre variable de vraies ruptures. Sur la figure de droite sont représentés les temps d'exécution sur les jeux de données contenant entre 100 et 1000 vraies ruptures.

Killick, R. and Fearnhead, P. and Eckley, I.A., Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, 107, 1590–98, 2012.

Lebarbier, E., Detecting multiple change-points in the mean of gaussian process by model selection, *Signal processing*, 85, 717–36, 2005.

Maidstone, R. and Hocking, T.D. and Rigaill, G. and Fearnhead, P., On optimal multiple changepoint algorithms for large data, *Statistics and computing*, 27, 519-33, 2017.

Picard, F. and Robin, S. and Lebarbier, E. and Daudin, J.J., A segmentation/clustering model for the analysis of array CGH sata, *Biometrics*, 63, 758–66, 2007.

Rigaill, G., A pruned dynamic programming algorithm to recover the best segmentations with 1 to k max change-points, *Journal de la Sociétée Française de Statistique*, 156, 180-205, 2015.

Verzelen, N. and Fromont, M. and Lerasle, M. and Reynaud-Bouret, P., Optimal Change-Point Detection and Localization, *arXiv:2010.11470*, 2020

textbfYao, Y.C. and Au, S.T., Least-squares estimation of a step function, *The Indian Journal of Statistics*, 51, 370-81, 1989.

Zhang, N.R. and Siegmund, D.O., A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data, *Biometrics*, 63, 22–32, 2007.

Une PLS parcimonieuse entre Statistique et Apprentissage

Hadrien Lorenzo [1] & Olivier Cloarec [2] & Rodolphe Thiébaut [3] & Jérôme Saracco [1]

[1] *ASTRAL, INRIA BSO, 200 Avenue de la Vieille Tour, 33405 Talence, France*
[2] *Research Group for Chemometrics, Institute of Chemistry, Umeå University, Umeå, S-901 87, Suède*
[3] *SISTM, INRIA BSO, 200 Avenue de la Vieille Tour, 33405 Talence, France*

Résumé. Ce travail s'intéresse à l'étude d'une méthode PLS (« Partial Least Squares » pour « Moindres Carrés Partiels ») parcimonieuse. Alors que les autres méthodologies de cette sorte travaillaient sur un modèle PLS déjà construit ou en construction, notre solution s'attaque à offrir une bonne estimation des matrices de covariance successives, qui sont au cœur de l'estimation du modèle PLS. Grâce à une régularisation par seuillage doux, le modèle PLS résultant est naturellement parcimonieux en covariables et en variables réponses. Cette méthodologie nécessite l'emploi d'une méthode de validation robuste fondée sur le bootstrap. De plus, de nouveaux critères de validation, sur la base du R^2 et du Q^2 ont été introduit afin de construire un modèle PLS moins sensible au sur-apprentissage, particulièrement néfaste si le nombre de covariables est important et le nombre d'observations est faible.

Mots-clés. PLS, parcimonie, sur-apprentissage, R^2, Q^2, grande dimension, $n \ll p$

Abstract. This work focuses on the study of a sparse PLS (« Partial Least Squares ») method. While other methodologies of this kind worked on a PLS model already built or under construction, our solution tackles the heart of the estimation problem by focusing on providing a good estimation of the successive covariance matrices, which are at the heart of the PLS model estimation. Thanks to a soft thresholding regularization, the resulting PLS model is naturally parsimonious in covariates and response variables. This methodology requires the use of a robust validation method to which the bootstrap is suitable. In addition, new validation criteria based on R^2 and Q^2 have been introduced in order to build a PLS model less sensitive to over-fitting, which is particularly harmful if the number of covariates is large and the number of observations is low.

Keywords. PLS, sparsity, over-fitting, R^2, Q^2, high-dimensional data, $n \ll p$

1 Introduction

La quantité de descripteurs, notée p, captant le comportement d'un nombre réduit d'observations, noté n, s'est enrichie très récemment et avec elle les méthodes permettant

de les analyser. On pense aux données de séquençage en génétique par exemple, où p peut valoir plusieurs dizaines de milliers pour peu, voire très peu, d'observations : n de l'ordre de quelques dizaines parfois. Ces données sont associées à des questions de recherche reformulées en critères d'optimisation que l'analyste doit résoudre. Les cas $n << p$ conduisent à des solutions non uniques qui se révèlent bien souvent instables, victimes de ce que l'on appelle le sur-apprentissage. Ces instabilités sont partiellement décelables en ayant recours à des méthodologies de ré-échantillonage telles que la validation croisée (CV pour cross-validation) ou le bootstrap, au centre de notre travail actuel. L'objectif des méthodes de régularisation est de trouver un équilibre entre cette stabilité des résultats et l'adéquation aux données, autrement dit la précision avec laquelle on respecte le problème d'optimisation ou encore, d'un point de vue plus pratique, la pertinence de la réponse à la question de recherche.

On suppose que $\mathbf{x}$ et $\mathbf{y}$ sont des variables aléatoires de dimensions p et q, $\mathbf{X}$ et $\mathbf{Y}$ sont des matrices aléatoires de dimension $n \times p$ et $n \times q$ où, pour chaque ligne $i \in [\![1, n]\!]$, $\mathbf{x}_i \sim \mathbf{x}$ et $\mathbf{y}_i \sim \mathbf{y}$ et les $\mathbf{x}_i$, respectivement les $\mathbf{y}_i$, sont indépendants les uns des autres. Les objets représentés par la lettre « y » doivent être prédits (réponse) par ceux représentés par la lettre « x » (prédicteurs). On note $\mathcal{D}_n$ le jeu de données basé sur ces n observations.

De nombreuses méthodes de régularisation ont émergé depuis une cinquantaine d'années et nous nous intéressons ici à la méthode des Moindres Carrés Partiels (PLS dans la suite pour « Partial Least Squares » en anglais) décrite initialement par Wold [3]. La PLS est une méthode de régression multi-linéaire capable d'estimer des matrices de régression même dans le cas $n < p$, où $\mathbf{X}'\mathbf{X}$ est non inversible. Le principe de cet algorithme est d'estimer itérativement, ici pour l'itération r, les premiers vecteurs singuliers droit et gauche, $\mathbf{u}_r$ et $\mathbf{v}_r$, des matrices de covariance empiriques $\mathbf{M}^{(r)} = \mathbf{Y}^{(r)\prime}\mathbf{X}^{(r)}/(n-1)$ où $\mathbf{Y}^{(r)}$ et $\mathbf{X}^{(r)}$ sont les matrices résiduelles des étapes précédentes $\mathbf{Y}^{(r-1)}$ et $\mathbf{X}^{(r-1)}$ sur le score $\mathbf{t}_{r-1} = \mathbf{X}^{(r-1)}\mathbf{u}_{r-1}$.

La suppression de variables inutiles, car non associées à la variable réponse, d'un modèle de prédiction induit une stabilisation du dit modèle en améliorant le rapport signal sur bruit. Assez naturellement, la problématique de suppression de ces variables inutiles est devenue un pan entier de recherche en mathématiques appliquées. On parle de modèles parcimonieux, « sparse » en anglais. Ce genre de méthodologie a été appliquée à l'analyse PLS au travers de nombreux travaux et [2] en décrit 16 versions au travers de 3 grandes classes nommées par les auteurs « filter », « wrapper » et « embedded » soit en français, respectivement, « filtrage », « enveloppement » et « encastrement » par exemple. Dans la première classe, un modèle PLS est filtré pour sélectionner les variables d'importance, bien souvent à partir d'un seuil unique. Les méthodes d'enveloppement reprennent une méthodologie de filtration pour réajuster un modèle PLS, ce qui est répété un certain nombre de fois. La troisième classe est occupée par des méthodes où la sélection de variables est réalisée en même temps que l'estimation des paramètres du modèle.

Nous proposons ici une autre vision de la sélection de variables en PLS qui recherche la stabilisation des matrices de covariance empiriques $\mathbf{M}^{(r)}$ en amont de chaque décomposi-

tion spectrale. La stabilisation est réalisée grâce à l'opérateur de seuillage doux appliqué directement à ces matrices et dont le fonctionnement pratique est expliqué plus loin. Cette méthodologie fait donc partie d'une quatrième classe de méthodes, en reprenant la classification de [2], qui tente de résoudre la sélection de variables avant l'estimation des paramètres. Elle travaille ainsi au plus près des données sans pour autant réaliser une sélection marginale univariée des prédicteurs. Cette proximité lui a valu la dénomination de « data driven » pour « motivée par les données » soit au total : « data driven sparse PLS » et dans la suite ddsPLS.
Afin que quantifier la qualité des modèle construits, nous nous sommes inspirés des notions de R^2 et de Q^2, coefficients très utilisés en analyse PLS et en régression de façon générale, pour construire une statistique, notée γ dans la suite, et des estimateurs associés. La sensibilité au phénomène de sur-apprentissage de ces estimateurs nous permet de sélectionner, parmi une famille de modèles ddsPLS concurrents, le modèle optimal. Ces estimateurs emploient des procédés de ré-échantillonnage bootstrap.

Dans la suite nous commencerons par décrire le modèle statistique des variables latentes puis le modèle ddsPLS. Ensuite nous détaillerons la statistique introduite ci-dessus et ses estimateurs. Une dernière section permet d'apprécier cette méthodologie sur un jeu de données simulées simple.

2 Modèle statistique à variables latentes

En plus de ce qui a été décrit en introduction sur les objets $\mathbf{x}$ et $\mathbf{y}$ et leurs équivalents populationnels, nous ajoutons que les composantes de $\mathbf{x}$ et $\mathbf{y}$ sont de variance unitaire. De plus, les méthodes telles que la PLS ou l'ACP sont souvent associés à des modèles à variables latentes $\mathcal{R}$-dimensionnelles, notées $\boldsymbol{\phi}$ dans cet énoncé et qui vérifient $\mathbb{E}\boldsymbol{\phi} = \mathbf{0}_{\mathcal{R}}$, $\text{var}(\boldsymbol{\phi}) = \mathbb{I}_{\mathcal{R}}$ et

$$\begin{cases} \mathbf{x} = \mathbf{A}'\boldsymbol{\phi} + \boldsymbol{\epsilon} \text{ où } \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_{\mathcal{R}}]' \in \mathcal{M}_{\mathcal{R}\times p}(\mathbb{R}) \text{ avec } \|\mathbf{a}_r\| \neq 0, \\ \mathbf{y} = \mathbf{D}'\boldsymbol{\phi} + \boldsymbol{\xi} \text{ où } \mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_{\mathcal{R}}]' \in \mathcal{M}_{\mathcal{R}\times q}(\mathbb{R}) \text{ avec } \|\mathbf{d}_r\| \neq 0, \\ \text{avec } \boldsymbol{\psi} = (\boldsymbol{\phi}', \boldsymbol{\epsilon}', \boldsymbol{\xi}')' \text{ et } \text{var}(\boldsymbol{\psi}) \text{ diagonale.} \end{cases} \tag{1}$$

Ce système conduit naturellement à considérer le modèle de régression

$$\mathbf{y} \;=\; \mathbf{B}'\mathbf{x} \;+\; \mathbf{e}, \tag{2}$$

où $\mathbf{B} \in \mathbb{R}^{p\times q}$ est une matrice non stochasique, $\mathbf{e} \in \mathbb{R}^q$ est une erreur résiduelle et $\mathbf{e}$ et $\mathbf{x}$ sont indépendants. Il vient que la matrice $\mathbf{B}$ satisfait

$$(\mathbf{AB})'\boldsymbol{\phi} \;=\; \mathbf{D}'\boldsymbol{\phi} \underset{\mathbb{E}\,\cdot\,\boldsymbol{\phi}'}{\Longrightarrow} \mathbf{AB} \;=\; \mathbf{D} \tag{3}$$

qui n'admet pas de solution unique dans le cas général.
On peut aussi remarquer que le Système (1) n'est pas identifiable, en effet pour une

matrice $\mathbf{R} \in \mathcal{GL}_{\mathcal{R}}(\mathbb{R})$ quelconque et en notant $\mathbf{t} = \mathbf{R}'\boldsymbol{\phi}$, $\mathbf{P} = \mathbf{R}^{-1}\mathbf{A}$ et $\mathbf{C} = \mathbf{R}^{-1}\mathbf{D}$, le Système (1) conduit au système de variable latentes plus général, et principalement utilisé dans la communauté

$$\left\{ \begin{array}{l} \mathbf{x} = \mathbf{P}'\mathbf{t} + \boldsymbol{\epsilon} \\ \mathbf{y} = \mathbf{C}'\mathbf{t} + \boldsymbol{\xi} \\ \text{avec } \mathrm{var}(\mathbf{t}', \boldsymbol{\epsilon}', \boldsymbol{\xi}')' \text{ diagonale} \end{array} \right. , \tag{4}$$

où $\mathbf{t}$ est le vecteur de scores et $\mathbf{P}$ et $\mathbf{C}$ sont appelés poids. Puisque ces poids ne sont pas identifiables, la méthode PLS cherche à construire le sous-espace engendré par $\boldsymbol{\phi}$ au travers d'une matrice $\mathbf{R}$ inconnue.

3 Une PLS parcimonieuse, ddsPLS

La méthode PLS a été décrite par de nombreux algorithmes, nous considérons ici l'algorithme Nipals. C'est un algorithme itératif avec R itérations.
La solution que nous proposons ici utilise une famille de R coefficients $\lambda^{(r)}$ pour seuiller les matrices de covariances empiriques successives après déflations. De plus, cet algorithme construit un estimateur $\widehat{\mathbf{B}}$ sparse à la fois en « x » et en « y ».

$$\left. \begin{array}{l} \text{ddsPLS} \\ \widehat{\mathbf{B}} = \mathbf{U}\left(\mathbf{P}'\mathbf{U}\right)^{-1}\mathbf{C} \\ \mathbf{X}^{(1)} = \mathbf{X} \\ \mathbf{Y}^{(1)} = \mathbf{Y} \end{array} \right\} , \forall r \in [\![1, R]\!] \quad \left\{ \begin{array}{l} \text{(a)} \left\{ \begin{array}{rcl} \mathbf{u}_r & = & \overrightarrow{\mathrm{RSV}}\left(S_{\lambda^{(r)}}\left(\mathbf{M}^{(r)}\right)\right), \\ \mathbf{v}_r & = & \overrightarrow{\mathrm{RSV}}\left(S_{\lambda^{(r)}}\left(\mathbf{M}^{(r)'}\right)\right), \end{array} \right. \\ \text{(b) } \mathbf{t}_r = \mathbf{X}^{(r)}\mathbf{u}_r, \\ \text{(c) } \mathbf{p}_r = \mathbf{X}^{(r)'}\mathbf{t}_r / \mathbf{t}_r'\mathbf{t}_r, \\ \text{(d)} \left\{ \begin{array}{rcl} \boldsymbol{\Pi}_r & = & \mathrm{diag}(\{\delta_{\neq 0}(\mathbf{v}_r)_j\}_{j\in[\![1,q]\!]}), \\ \mathbf{c}_r & = & \arg\min_{\mathbf{V}} \left\|\mathbf{Y}^{(r)}\boldsymbol{\Pi}_r - \mathbf{t}_r\mathbf{V}'\right\|^2, \end{array} \right. \\ \text{(e) } \mathbf{X}^{(r+1)} = \mathbf{X}^{(r)} - \mathbf{t}_r\mathbf{p}_r', \ \mathbf{Y}^{(r+1)} = \mathbf{Y}^{(r)} - \mathbf{t}_r\mathbf{c}_r', \end{array} \right. \tag{5}$$

$\overrightarrow{\mathrm{RSV}}$ (pour Right-Singular-Vector) renvoie le 1^{er} vecteur singulier droit de son argument. L'opérateur de seuillage doux « $S_\lambda(\cdot)$ » retire à chaque élément de son argument matriciel une valeur $\lambda \geq 0$ et retourne 0 si le coefficient initial est inférieur (en valeur absolue) à λ. Ici (a) fixe les poids, (b) fixe les scores, (c) estime la matrice de régression de $\mathbf{X}^{(r)}$ sur $\mathbf{t}_r$ et (d) de $\mathbf{Y}^{(r)}$ sur $\mathbf{t}_r$ et (e) réalise la déflation de chacune des matrices.
Si $\lambda^{(r)} = 0$ alors on retrouve l'algorithme Nipals classique.

4 Une statistique et des estimateurs bootstrap

Le coefficient R^2, également appelée variance expliquée ou coefficient de détermination, décrit la qualité de l'ajustement d'un modèle prédictif. Il est défini par

$$R^2 = 1 - \frac{\sum_{j=1}^{q}\sum_{i=1}^{n}\left(y_{i,j} - \hat{y}_{i,j}\right)^2}{\sum_{j=1}^{q}\sum_{i=1}^{n}\left(y_{i,j} - \bar{y}_j\right)^2}, \tag{6}$$

où $\bar{y}_j$ est la moyenne empirique de y_j (composante j de $\mathbf{y}$) et $\hat{y}_{i,j}$ est l'estimation de $y_{i,j}$ par le modèle courant. On appelle « Somme des Carrés Résiduels » (RSS pour « Residual Sum of Squares ») le numérateur dans l'expression précédente et « Somme des Carrés Totaux » (TSS pour « Total Sum of Squares ») le dénominateur si bien que $R^2 = 1 - \text{RSS}/\text{TSS}$. Cette métrique rapporte donc l'inertie des erreurs en entraînement à l'inertie totale du jeu d'entraînement. Plus ce R^2 est grand et plus le modèle est proche de la structure du jeu de données d'entraînement et vaut 1 lorsque l'erreur est nulle.
Introduisons la statistique suivante

$$\gamma \quad = \quad 1 - \frac{\sum_{j=1}^{q} \text{var}(y_j - \hat{y}_j^{(\mathbf{p})})}{\sum_{j=1}^{q} \text{var}(y_j)}, \tag{7}$$

avec $\hat{y}_j^{(\mathbf{p})}$ l'estimateur de y_j via le modèle $\mathbf{p}$. On peut remarquer que l'expression du R^2, détaillée précédemment, est une estimation de γ pour le modèle $\mathbf{p}_n$ (construit sur $\mathcal{D}_n$) les erreurs étant estimées aussi sur $\mathcal{D}_n$. Il est ainsi possible d'interpréter la valeur $\gamma = 0$, qui correspond à un modèle prédictif de qualité équivalente à la prédiction à la moyenne. Si $\gamma > 0$, alors le modèle se trompe généralement moins que la moyenne, et sinon il se trompe plus. Il vient que si $\gamma > 0$, alors le modèle courant peut être conservé. Cet estimateur est sensible au sur-apprentissage notamment car les erreurs sont estimées sur le jeu de données qui a servi à construire le modèle. Afin de s'en affranchir, du moins partiellement, le Q^2 a été introduit, il se définit comme

$$Q^2 \quad = \quad 1 - \frac{\sum_{j=1}^{q} \sum_{f=1}^{F} \sum_{i \notin \text{cv}_f} \left(y_{i,j} - \hat{y}_{i,j}^{(\mathbf{p}_f)} \right)^2}{\text{TSS}}, \tag{8}$$

où F est le nombre de sous-échantillons (« fold » en anglais) obtenus par validation croisée et cv_f la liste des indices ne faisant pas partie du fold f. De même, $\hat{y}_{i,j}^{(\mathbf{p}_f)}$ est l'estimation de $y_{i,j}$ via le modèle $\mathbf{p}_f$.

Notre méthodologie nécessitant un lissage prononcé des courbes de validation, nous nous sommes intéressés à des estimateurs bootstrap. En effet, nous pouvons construire des estimateurs de γ en utilisant des échantillons bootstraps et une formulation plus riche du R^2 et du Q^2 : $\bar{R}^2_B = \frac{1}{B} \sum_{b=1}^{B} R^2_b$ et $\bar{Q}^2_B = \frac{1}{B} \sum_{b=1}^{B} Q^2_b$ comme moyennes empiriques des B estimateurs suivants

$$R^2_b \quad = \quad 1 - \frac{\sum_{j=1}^{q} \sum_{i \in \text{IN}(b)} \left(y_{i,j} - \hat{y}_{i,j}^{(\mathbf{p}_b)} \right)^2}{\sum_{j=1}^{q} \sum_{i \in \text{IN}(b)} \left(y_{i,j} - \bar{y}_j^{(b)} \right)^2}, \qquad Q^2_b \quad = \quad 1 - \frac{\sum_{j=1}^{q} \sum_{i \in \text{OOB}(b)} \left(y_{i,j} - \hat{y}_{i,j}^{(\mathbf{p}_b)} \right)^2}{\sum_{j=1}^{q} \sum_{i \in \text{OOB}(b)} \left(y_{i,j} - \bar{y}_j^{(b)} \right)^2}, \tag{9}$$

où $\text{IN}(b)$ et $\text{OOB}(b)$ sont les indices des observations In-Bag (« dans le sac », IB) et Out-Of-Bag (« en dehors du sac », OOB). $\hat{y}_{i,j}^{(\mathbf{p}_b)}$ est l'estimation de $y_{i,j}$ par le modèle $\mathbf{p}_b$ et $\bar{y}_j^{(b)}$ est la moyenne empirique de y_j calculée sur l'échantillon bootstrap b.

En analyse PLS, il est courant de conserver le modèle qui minimise l'écart entre le R^2 et le Q^2, voir [1]. En effet, les deux descripteurs sont en réalité sujets au sur-apprentissage et le maximum de Q^2 conduit bien souvent à des modèles non optimaux. Nous utilisons alors la métrique « $\bar{R}^2_B - \bar{Q}^2_B$ » pour sélectionner le modèle optimal avec la condition « $\bar{Q}^2_B > 0$ » pour pouvoir conserver le modèle.

5 Exemple sur un jeu de données simple

Considérons la structure de données suivante

$$\mathbf{A} = \sqrt{1-\sigma^2}\begin{pmatrix} \mathbf{1}'_{50} & \mathbf{0}'_{950} \end{pmatrix}, \quad \mathbf{D} = \sqrt{1-\sigma^2}\begin{pmatrix} 1 \end{pmatrix}, \tag{10}$$

avec $1-\sigma^2 = 0.9025$ où σ est tel que

$$\boldsymbol{\psi} = (\phi, \boldsymbol{\epsilon}_{1\ldots50}/\sigma, \boldsymbol{\epsilon}_{51\ldots1000}, \xi/\sigma) \sim \mathcal{N}\left(\mathbf{0}, \mathbb{I}_{1+1000+1}\right).$$

Dans cette structure, les 50 premiers coefficients de la matrice de régression (qui est ici un vecteur) devraient être égaux et différents de 0 alors que les 950 derniers devraient être égaux à 0. Afin d'illustrer le comportement des estimateurs $\bar{R}^2_B$ et $\bar{Q}^2_B$ de γ, on simule trois jeux de données pour trois tailles d'échantillons ($n = 50, 100, 200$) que l'on soumet au bootstrap pour le modèle ddsPLS. On trace alors les courbes associées pour des valeurs de λ comprises entre 0 et 1, voir la première ligne de la Figure 1. La première figure de la seconde ligne représente le nombre de variables sélectionnées pour chaque modèle et la seconde les boxplots des 50 premiers coefficients de régression, les autres étant égaux à 0. Ces modèles sont construits sur une composante, par convergence de l'algorithme ddsPLS.

On observe que les $\bar{R}^2_B$ sont généralement importants là où les $\bar{Q}^2_B$ sont généralement faibles, ceci pour des zones de faibles régularisations (λ faibles et beaucoup de variables sélectionnées) ou inversement pour des zones de fortes régularisations (λ important et trop peu de variables sélectionnées). Ainsi les $\bar{Q}^2_B$, dans les trois expériences, montrent des zones raisonnables pour λ entre 0.3 et 0.8, au travers d'un plateau qui correspond à exactement 50 variables sélectionnées. Pourtant le choix précis de λ sur ce genre de plateau est complexe. En effet, quel que soit n, le maximum en $\bar{Q}^2_B$ (représenté par le symbole « ✡ ») correspond à une zone de sur-apprentissage puisque le $\bar{R}^2_B$ commence à augmenter et que le nombre de variables sélectionnées est supérieur à 50 (visible ici pour $n = 100$). Les points sélectionnées, qui correspondent aux minima en « $\bar{R}^2_B - \bar{Q}^2_B$ », se positionnent plutôt en milieu de plateau et permettent d'éviter ainsi le sur-apprentissage.

6 Conclusion

La méthode présentée dans ce travail propose une solution afin de gérer les effets délétères de la grande dimension par régularisation parcimonieuse. Lorsque $\mathbf{y}$ est multi-

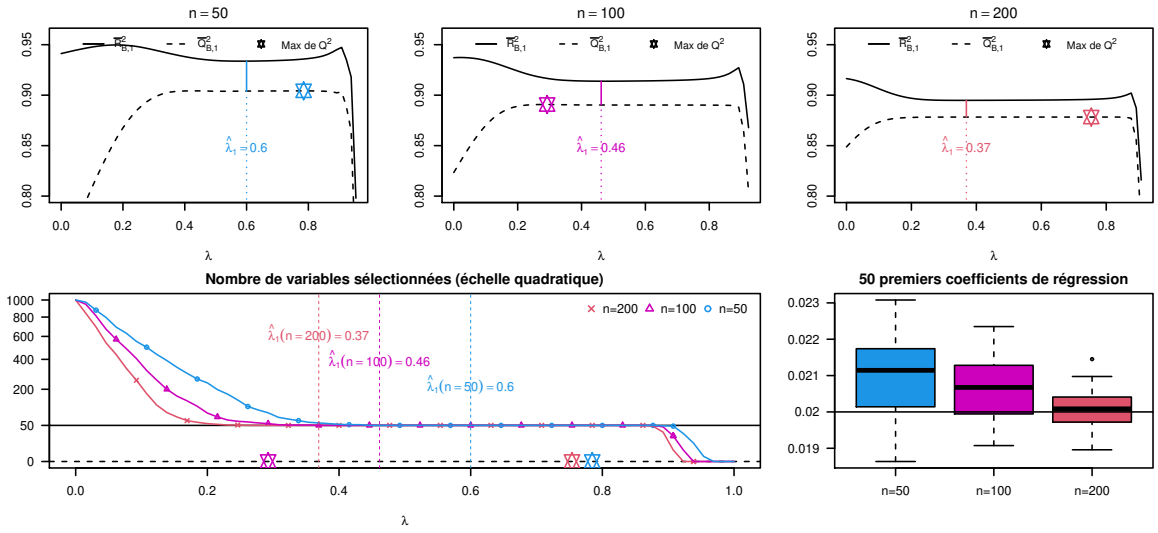


FIGURE 1 – Résultats de simulations pour différents n. Dans le dernier graphique on ne fait pas apparaître les 950 derniers coefficients qui sont égaux à 0 quel que soit n.

dimensionnel, cette méthodologie montre des résultats très encourageants (non exposés ici). De nouveaux critères statistiques ont été explorés, basés sur des critères existants et reconnus, utilisant avantageusement les effets de sur-apprentissage intrinsèquement associés aux descripteurs R^2 et Q^2. Là où la Statistique suit la métrique du R^2, l'Apprentissage suit le Q^2 et notre méthodologie se nourrit de la différence de ces deux critères.
Des simulations ont permis de comparer cette méthodologie à d'autres déjà établies et reconnues. Leurs discussions feront l'objet de futurs travaux.

Références

[1] Olivier CLOAREC. « Can we beat over–fitting ? » In : *Journal of Chemometrics* 28.8 (2014), p. 610-614.

[2] Tahir MEHMOOD, Solve SÆBØ et Kristian Hovde LILAND. « Comparison of variable selection methods in partial least squares regression ». In : *Journal of Chemometrics* 34.6 (2020). e3226 cem.3226, e3226. DOI : https://doi.org/10.1002/cem.3226. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.3226. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.3226.

[3] Herman WOLD. « Estimation of principal components and related models by iterative least squares ». In : *Multivariate analysis* (1966), p. 391-420.

Détection de ruptures faibles dans la moyenne des modèles CHARN

Marwa Ltaifa [1] & Joseph Ngatchou-Wandji [2]

[1] *IECL, Université de Lorraine, France & LAMMDA, Université de Sousse, Tunisie.*
E-mail : marwa.ltaifa@univ-lorraine.fr
[2] *IECL,Université de Lorraine, France.*
E-mail : joseph.ngatchou-wandji@univ-lorraine.fr

Résumé. Dans [7] sont présentées des méthodes et stratégies pour détecter et estimer les localisations des ruptures dans la moyenne d'une grande classe de modèle autoregressifs conditionnellement non-linéaire. Nous faisons des simulations pour illustrer ces méthodes et les appliquer à la détection des ruptures dans des données du Covid-19.
Mots-clés. Séries chronologiques, Ruptures, Données du Covid-19 en France.

Abstract. In [7] is presented some methods and strategies for detectiong and estimating the locations of weak changes in the mean of classe of a Conditional Heteroscedastic AutoRegressif Non-linear model (CHARN) models. We present the results of some simulations for illustrating these methods which are apply to detecting changes in Covid-19 data.
Keywords. Time series, Breaks, Covid-19 data in France.

1 Introduction

Dans le présent travail on s'intéresse à l'étude des petites ruptures dans la moyenne des modèles CHARN. En santé, celles-ci peuvent être des signaux annonciateurs de maladies. En finance, elles peuvent annoncer une crise financières. En climatologie, elles peuvent signaler une tempête, une sécheresse, une inondation ou encore une canicule.
Des nouvelles méthodes sont proposées dans [7] pour détecter ce type de ruptures, ainsi que des stratégies pour estimer leurs localisations. Nous les appliquons à la détection des ruptures dans la moyenne des données sur le nombre de décès quotidiens du Covid-19 en France lors de la première vague. Au préalable, quelques résultats de simulation sur des données issues d'un modèle CHARN sont présentées.

2 Les méthodes

Les méthodes étudiées dans [7] reposent essentiellement sur la puissance théorique d'un test du rapport de vraisemblance pour discriminer entre modèles conditionnellement

hétéroscédastique non-linéaires CHARN). Plus précisément, soit une série d'observations $X_1, X_2, \ldots, X_n$ générée par le modèle CHARN(p) suivant

$$X_t = T(Z_{t-1}) + \gamma^\top \omega(t) + V(Z_{t-1})\varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

où $\gamma = (\gamma_1, \ldots, \gamma_k, \gamma_{k+1})^\top \in \mathbb{R}^{k+1}$ et pour $t_1, \ldots, t_k$, $1 < t_1 < \ldots < t_k < n$,
$\omega(t) = (\mathbb{1}_{[1,t_1[}(t), \mathbb{1}_{[t_1,t_2[}(t), \ldots, \mathbb{1}_{[t_k,n[}(t))^\top \in \{0,1\}^{k+1}$, $(X_t)_{t\in\mathbb{Z}}$ est un processus stationnaire par morceaux et ergodique, $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc centré réduit de densité f, pour tout $t \in \mathbb{Z}$, $Z_t = (X_t, \ldots, X_{t-p+1})^\top$, $p \in \mathbb{N}$, et T et V sont des fonctions réelles telles que $\inf_{x\in\mathbb{R}^p} V(x) > 0$. Parmi les travaux qui ont étudié cette classe de modèles nous pouvons citer par exemple [1], [2], [3] et [6].
Dans [7] un test du rapport de vraisemblance est construit pour tester

$$H_0 : \gamma = \gamma_0 \text{ contre } H_\beta^{(n)} : \gamma = \gamma_0 + \frac{\beta}{\sqrt{n}} = \gamma_n, \quad n > 1,$$

pour $\gamma_0 \in \mathbb{R}^{k+1}$ et $\beta \in \mathbb{R}^{k+1}$. Ces deux hypothèses se rapprochent lorsque la taille de l'échantillon grandit. On montre qu'elles sont contiguës au sens de Le Cam (voir [4]). Cette propriété permet l'étude de la puissance du test construit, et l'obtention d'une expression explicite de sa puissance. En effet, si $\psi_0 = (\rho_0^\top, \theta_0^\top)^\top \in \Theta \times \widetilde{\Theta} \subset \mathbb{R}^l \times \mathbb{R}^q$ est le vrai paramètre de nuisance du modèle (1), sous certaines hypothèses techniques, on montre que pour toute valeur de β, le test du rapport de vraisemblance construit est asymptotiquement optimal, de puissance asymptotique locale $\mathcal{P}_{k,t^k} = 1 - \Phi\left(u_\alpha - \varpi(\gamma_0, \beta)\right)$, où Φ est la fonction de répartition de la loi normale centrée réduite, u_α le quantile d'ordre $1-\alpha$, $\alpha \in (0,1)$ et ω est une fonction à valeurs réelles dont nous ne rappelons pas l'expression qui se trouve dans [7].
La première étape des méthodes décrites dans [7] consiste à déterminer, à partir du chronogramme, les m premières données $X_1, X_2, \ldots, X_m$ qui sont à peu près stationnaires, le nombre maximum K de ruptures potentielles et la distance minimale $h << n$ entre elles. Notons $\mathcal{P}_{0,t^0} = \alpha$, et considérons $\tau \in (0,1)$. Pour détecter la présence de rupture, prendre $k = 1$ et appliquer le test à tous les t_1 tels que $m \le t_1 \le n - h$.

1. Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| \le \tau$ pour tous ces t_1, alors aucune rupture n'est détectée dans la série.
2. Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| > \tau$ pour un t_1, alors, il existe au moins une rupture dans la série.

Pour estimer les localisations des ruptures, pour $k = 1, \ldots, K$, on suppose que $m < \tau_1^0 < \ldots < \tau_k^0 < n - h$, $\tau_j^0 - \tau_{j-1}^0 \ge h$, $j = 2, \ldots, k$, sont de potentielles localisations des ruptures obtenues du chronogramme. Soit C_j un ensemble arbitraire d'indices autour des τ_j^0, $j = 1, \ldots, k$. On considère $S_k = C_1 \times C_2 \times \ldots C_k$. Pour tout k-uplet $\tau^k = (\tau_1, \ldots, \tau_k) \in S_k$, on applique le problème de test ci-dessus avec $t_j = \tau_j$, $j = 0, \ldots, k+1$ et on calcule $\mathcal{P}_{k,t^k}$.
• À l'étape $k+1$:

1. Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \mathcal{P}_{k,t^k}| \leq \tau$ et $|\widehat{\mathcal{P}}_{k,t^k} - \mathcal{P}_{k-1,t^{k-1}}| > \tau$, alors nous estimons le couple $(\widehat{k}, \widehat{t}^k)$ du nombre des ruptures et du vecteur des localisations par

$$(\widehat{k}, \widehat{t}^k) = \arg\max_{t^k \in S_k} \widehat{\mathcal{P}}_{k,t^k}.$$

2. Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \mathcal{P}_{k,t^k}| > \tau$, nous répétons l'étape 1 en remplaçant k par $k+1$.

3 Ruptures et estimation des localisations

Dans ce paragraphe, nous utilisons le logiciel R pour étudier les performances des résultats théoriques obtenus dans [7]. Nous appliquons ces résultats au modèle (1) pour $p = 1$, $Z_t = X_{t-1}$, $T(Z_{t-1}) = \rho_1 + \rho_2 X_{t-1} e^{\rho_3 X_{t-1}^2}$, $\gamma = \gamma_0 + \beta/\sqrt{n}$ et $V(Z_{t-1}) = (\theta_1 + \theta_2 X_{t-1}^2 e^{-\theta_3 X_{t-1}^2})^{1/2}$, où les ρ_j, θ_j et γ_0 sont des paramètres prenant certaines valeurs à préciser, n est la taille de l'échantillon, $(\varepsilon_t)_t$ est un bruit blanc standard de densité f, et β est un réel arbitraire. Le niveau nominal considéré est $\alpha = 0.05$ et le nombre de réplications est $N = 5000$.

3.1 Simulations

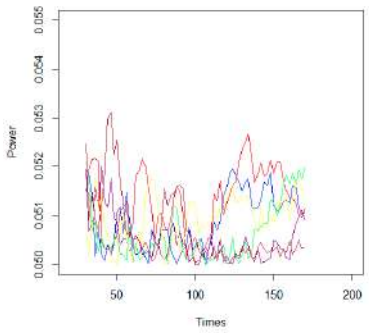
Plusieurs situations sont considérées pour évaluer la performance de notre méthode. Nous commençons par le cas d'une série d'observations stationnaires. C'est-à-dire une série sans ruptures. Nous considérons ensuite le cas d'une série comportant une rupture.

3.1.1 Aucun point de rupture dans les données

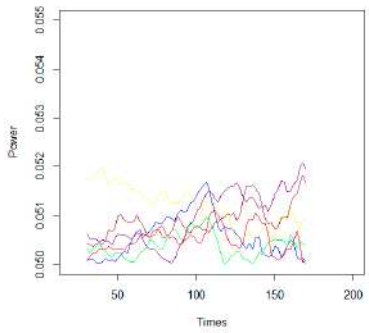
Nous calculons d'abord la puissance locale asymptotique dans le cas où le modèle ne présente aucune rupture, c'est-à-dire dans le cas où $k = 0$. Nous considérons le modèle ci-dessus lorsque $n = 200$, $\gamma_0 = 0$ et f la densité gaussienne standard. Puis, nous représentons graphiquement la puissance dans le cas où $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$ et dans le cas où $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$. Les deux courbes sont construites sur la Figures 1.
On observe que la puissance du test ne dépasse pas 0.053 pour $\alpha = 0.05$. Donc on accepte l'hypothèse nulle d'absence de rupture, ce qui est bien le cas ici.

3.1.2 Un seul point de rupture

Nous prenons le cas où $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$ et le cas où $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$. Pour chaque cas considéré, nous traçons les courbes de la puissance du test en faisant varier les instants de rupture (voir les Figures 2 (a) et 2 (b)). Par exemple pour le premier échantillon considéré, nous prenons $t_1 = 30$ c'est-à-dire $n_1(n) = 30$ et $n_2(n) = 170$ et nous traçons la puissance du test pour $\beta = (0; -0.5)$ et pour $\beta = (0; 0.8)$. Ensuite, nous prenons $t_1 = 60$, c'est-à-dire $n_1(n) = 60$ et $n_2(n) = 140$ et nous traçons la courbe de la puissance pour $\beta = (0; 0.4)$ et pour $\beta = (0; -0.4)$ ainsi de suite. On constate

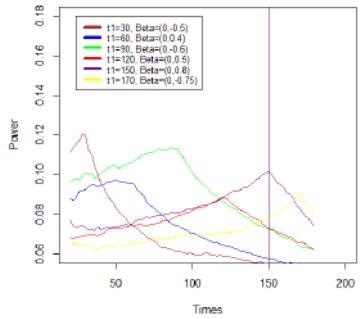


(a) $n = 200, \gamma_0 = 0, \rho_1 = \rho_2 = 0, \theta_1 = 1, \theta_2 = 0$

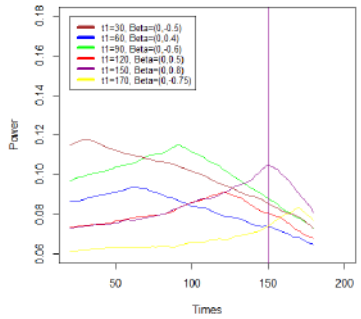


(b) $n = 200, \gamma_0 = 0, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1, \theta_2 = 0$

FIGURE 1 – Pas de rupture



(a) $n = 200$, $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$



(b) $n = 200$, $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$

FIGURE 2 – Une rupture

que la puissance du test est maximale à l'instant de rupture. Donc notre test détecte bien les instants de rupture qui sont les temps donnant la plus grande puissance.

3.2 Application aux données réelles

Nous cherchons dans cette partie les points de rupture dans la série sur le décès quotidiens de COVID-19 en France dans la période du 27/02/2020 au 10/07/2020. La Figure 3 (a) correspond au chronogramme des données brutes. Nous voulons savoir si les potentiels points de rupture dans cette série, tracés en vert dans la figure, représentent réellement des points de rupture. Ces points sont $t_1 = 31, t_2 = 38, t_3 = 55, t_4 = 86$ et $t_5 = 116$ correspondants respectivement aux dates suivantes : 28/03/2020, 04/04/2020, 21/04/2020,

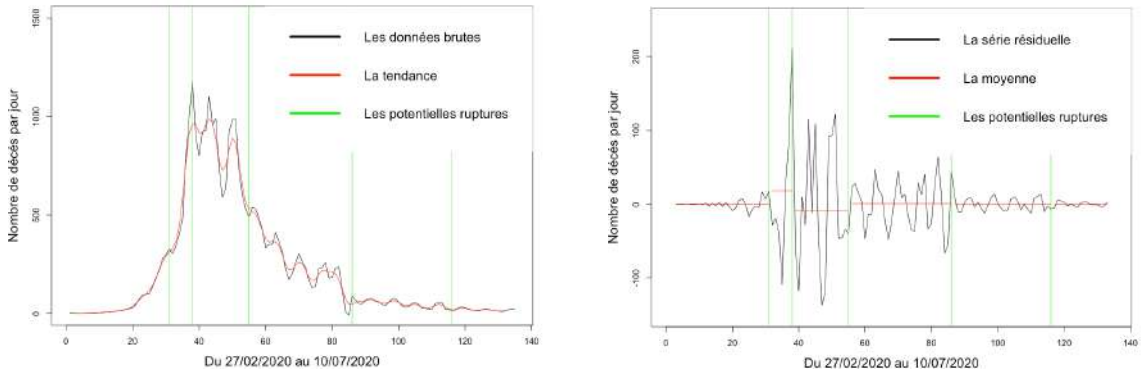
22/05/2020 et 21/06/2020. Nous commençons tout d'abord par modéliser la série que nous étudions.
D'après le graphique, celle-ci présente une tendance et ne semble pas présenter de saisonnalité. Les résultats de [7] ne peuvent pas s'appliquer directement à ces séries. Nous considérons alors la série corrigée de la tendance par la méthode des moyennes mobiles d'ordre 5. Cette série est représentée dans la Figure 3. (b). Sur chaque intervalle $[t_{i-1}, t_i)$ la série résiduelle semble stationnaire. Nous ajustons à cette série un modèle de la forme

$$X_t = \mu + (\beta_i/\sqrt{n}) + \sigma_i \varepsilon_t,$$

où pour tout $i \in \{1, \ldots, 6\}$, $\mu + (\beta_i/\sqrt{n})$ et σ_i sont respectivement la moyenne et la variance de X_t sur l'intervalle $[t_{i-1}, t_i)$, $\beta_1 = 0$ et pour tout $i = 2, \ldots, 6$, $\beta_i \in \mathbb{R}$, ε_t est un bruit blanc gaussien centré réduit, $t_0 = 1$, $t_6 = n$ et 1 et n sont respectivement 27/02/2020 et 10/07/2020. Cela correspond à notre problème de test pour $T(x) = \mu$, $V(x) = \sigma_i$ sur chaque intervalle $[t_{i-1}, t_i)$, $\gamma_0 = (\mu, \mu, \ldots, \mu)^\top$, $\gamma_n = (\mu, \mu + (\beta_2/\sqrt{n}), \mu + (\beta_3/\sqrt{n}), \ldots, \mu + (\beta_6/\sqrt{n}))^\top$ et $\beta = (0, \beta_2, \beta_3, \ldots, \beta_6)^\top$. Nous calculons la puissance du test autour des t_i et nous prenons les dates donnant la plus grande puissance. Nous obtenons ainsi les dates : $\widehat{t_1} = 35, \widehat{t_2} = 40, \widehat{t_3} = 56, \widehat{t_4} = 88$ et $\widehat{t_5} = 1117$, à compter à partir du 27/02/2020. Ces dates correspondent respectivement aux dates 02/03/2020, 06/04/2020, 22/04/2020, 21/05/2020 et 25/06/2020, différentes mais assez proches des dates potentielles de rupture. Elles sont représentées dans la Figure 4.
Une interprétation possible des dates estimées est la suivante : Au tour du 02/03/2020, le nombre de décès augmente drastiquement, atteint son pic et redescend autour du 06/04/2020, puis oscille significativement jusqu'aux environs du 22/04/2020, et un peu moins significativement entre les première et deuxième phases du déconfinement, qui ont lieu le 11/05/2020 et le 02/06/2020 respectivement, jusqu'au 21/05/2020, date à partir laquelle il se stabilise avant de se réduire considérablement à partir du 25/06/2020, peu après la troisième phase du déconfinement, qui a lieu le 02/06/2020.

3.3 Comparaison avec d'autres méthodes

D'autres méthodes issues pour le CUSUM test sont étudiées dans [5] dans le but est de détecter les ruptures lorsqu'elle se produit dans les premières ou les dernières observations. Nous voulons dans ce paragraphe comparer les résultats obtenues par nos méthodes à ceux obtenues par [5] implémentées sous R. Pour cela, nous simulons une série d'observations générées par le modèle (1) pour $n = 200, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1$ et $\theta_0 = 0$. Nous prenons le cas où on a qu'un seul point de rupture à l'instant $t = 30, 60, 90, 120, 150, 170$ et le paramètre $\beta = (0, \beta_2)$, où β_2 est une valeur arbitraire dans $\mathbb{R}$. Nous notons notre méthode NEW et les deux méthodes de [5] par SCUSUM et RCUSUM. Les résultats sont affichés dans la Table 1.
Nous remarquons que lorsque β_2 prend de petites valeurs, les SCUSUM et RCUSUM donnent de mauvais résultats. Ces deux méthodes n'estiment pas bien les localisations



(a) La série des données brutes (b) La série résiduelle

FIGURE 3 – Nombre de décès quotidiens de COVID-19 en France du 27/02/2020 au 10/07/2020.

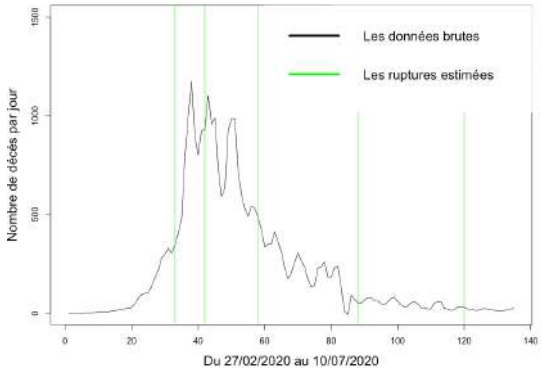


FIGURE 4 – Les dates de rupture trouvées

des points de rupture. Si β_2 prend de grandes valeurs, les deux méthodes donnent des estimations plus proches de l'instant exact de rupture, mais RCUSUM est plus performante SCUSUM. Dans tous les cas, il est clair que la méthode NEW est la plus efficace. Elle estime bien la localisation des points de rupture et donne des résultats plus proches que les deux autres méthodes, quelle que soit la valeur de β_2.
De plus, lorsque nous prenons $\beta = (0, 0)$ c'est-à-dire lorsque la série ne présente aucune rupture (lorsqu'elle est stationnaire), les SCUSUM et RCUSUM estiment des localisations des points de rupture qui n'existent pas. La méthode NEW est donc bien plus efficace et plus performante que les deux autres sur l'exemple considéré.

Méthodes	t exact $(0, \beta_2)$	30 $(0, -0.5)$	60 $(0, 0.4)$	90 $(0, -0.6)$	120 $(0, 0.5)$	150 $(0, 0.8)$	170 $(0, -0.75)$
NEW		30	60	91	120	149	170
SCUSUM		99	99	99	100	101	100
RCUSUM		91	60	69	143	79	126
Méthodes	t exact $(0, \beta_2)$	30 $(0, -5)$	60 $(0, 4)$	90 $(0, -6)$	120 $(0, 5)$	150 $(0, 8)$	170 $(0, -2)$
NEW		30	60	90	120	150	171
SCUSUM		76	81	93	113	136	105
RCUSUM		29	63	82	80	159	44
Méthodes	t exact $(0, \beta_2)$	30 $(0, -25)$	60 $(0, 24)$	90 $(0, -26)$	120 $(0, 25)$	150 $(0, 28)$	170 $(0, -22)$
NEW		30	60	90	120	150	170
SCUSUM		36	62	91	120	148	163
RCUSUM		38	61	90	121	150	166

TABLE 1 – $n = 200, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1$ et $\theta_0 = 0$.

Références

[1] Amano, T. (2012). *Asymptotic Optimality of Estimating Function Estimator for CHARN Model*. Advances in Decision Sciences.

[2] Bardet, J. M., & Kengne, W. (2014). *Monitoring procedure for parameter change in causal time series*. Journal of Multivariate Analysis, 125, 204-221.

[3] Bardet, J. M., & Wintenberger, O. (2009) *Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes*. The Annals of Statistics, 37(5B), 2730-2759.

[4] Dreosebeke, J-J. & Fine, Inférence non paramétrique : Les statistiques de rangs. (1996). Ed. de l'Université de Bruxelles ; Ed. Ellipses.

[5] Horváth, L., Miller, C., & Rice, G. (2020). *A new class of change point test statistics of Rényi type*. Journal of Business & Economic Statistics, 38(3), 570-579.

[6] Kengne, W. C. (2012). *Testing for parameter constancy in general causal time-series models*. Journal of Time Series Analysis, 33(3), 503-518.

[7] Ngatchou-Wandji, J., & Ltaifa, M. (2021). *On detecting weak changes in the mean of CHARN models*. arXiv preprint arXiv :2101.08597.

Apprentissage par renforcement pour les enchères en temps réel

Slimane Makhlouf [1] & Avner Bar-Hen [2] & François-Xavier Jollois [3]

[1] *smakhlouf@velvetconsulting.com*
[2] *avner@cnam.fr*
[3] *francois-xavier.jollois@u-paris.fr*

Résumé. Les enchères en temps réel (Real-time bidding, RTB) revêtent depuis la dernière décennie une grande importance pour les annonceurs de publicité en ligne. Cet intérêt grandissant conduit les acteurs à développer des techniques de plus en plus sophistiquées. Le travail que nous présentons se focalisent sur l'étude et la formalisation de ce problème sous forme d'un processus de décision markovien, ainsi que le développement d'algorithmes d'enchère efficaces et compétitifs. Nous utilisons le dataset public Ipinyou qui est très largement utilisé dans tous les travaux similaires et qui nous permet de comparer les performances de nos algorithmes.

Mots-clés. Enchères en temps réel, Apprentissage automatique, Apprentissage par renforcement, Processus de décision Markovien, algorithme acteur-critique

Abstract. Real-time bidding (RTB) has been of great importance to online advertisers for the last decade. This growing interest has led the players to develop increasingly sophisticated techniques. The work we present focuses on the study and formalization of this problem in the form of a Markovian decision process, as well as the development of efficient and competitive bidding algorithms. In our experiments, we use the Ipinyou dataset which is largely used in similar research work enabling us to compare our algorithms' performances.

Mots-clés. Real-time bidding, Machine learning, Reinforcement learning, Markov decision process, actor-critic algorithm.

1 Introduction

Le « real-time bidding » (RTB, enchère en temps réel) consiste à vendre en temps réel et au plus offrant un espace publicitaire sur une page web. Le but d'un annonceur est de remporter, pour un coût minimal, les enchères dont il espère un clic (puis un achat). Le prix de cette enchère est basé en général sur son historique d'enchères, de la potentialité des publicités à donner lieu à un clic, de son budget ainsi que de caractéristiques de l'internaute ayant ouvert la page contenant l'espace publicitaire mis aux enchères.
Nos travaux se focalisent sur ce point de vue « annonceur » et explorent les techniques permettant la maximisation du revenu (nombre de clics) sous différentes contraintes inhérentes à l'écosystème du RTB que nous détaillons dans la suite.

2 Processus de décision markovien et RTB

2.1 Formulation du RTB sous forme de processus de décision Markovien

Le problème du RTB se prête très bien à une modélisation sous forme de processus de décision markovien (MDP) [3]. Formellement, Un MDP est défini :

- S : L'ensemble des états possibles de l'environnement du point de vue de l'annonceur.
- A : L'ensemble des actions offertes à l'annonceur afin d'interagir avec l'environnement.
- T : La matrice de transition $T : S \times A \to S$ qui représente les dynamiques de l'environnement : $T(s, a, s')$ est la probabilité d'arriver dans l'état s' en prenant l'action a depuis l'état s.
- R : La matrice de récompense $R : S \times A \times S \to \mathbb{R} : R(s, a, s')$ qui représente la récompense associée à l'action a prise depuis l'état s menant à l'état s'.

L'annonceur est chargé de prendre des actions, ici des enchères, en considérant l'état actuel de la campagne d'affichage publicitaire (budget restant, nombre d'enchères restants, etc.. . .). L'environnement du MDP correspond à la salle des enchères : il est chargée d'adjuger les enchères, de faire payer les gagnants mais aussi de distribuer les éventuelles récompenses que sont les clics.

Lors de la connexion d'un utilisateur à une page ayant un espace publicitaire disponible, une requête d'affichage contenant les informations de connexions de l'utilisateur (navigateur et appareil utilisé, région, IP, . . .) ainsi que des informations sur l'encart publicitaire (taille, position, . . .) est envoyée et diffusée sur une plate-forme (appelée AdExchange), et qui correspond à l'environnement. Les annonceurs qui souhaitent participer à l'enchère au temps t formulent ainsi leur proposition de prix $\mathbf{b_t}$. L'enchère est remportée par le plus offrant et adjugée au prix de la deuxième enchère la plus élevée (Second price auction). Enfin la publicité est transmise et affichée.

Afin de calculer le meilleur prix ou la meilleure action du point de vue du MDP, chaque annonceur doit en premier lieu évaluer la probabilité $\mathbf{p}(\mathbf{x})$ que l'affichage de sa publicité entraîne une action (clic, achat, . . .) de la part de l'utilisateur qui voit la publicité. La probabilité $\mathbf{p}(\mathbf{x})$ d'un clic est un paramètre important pour optimiser une campagne d'affichage en ligne. En pratique, le taux de clic (τ) moyen est souvent inférieur à 0.1% et est estimé par chaque annonceur sur ses campagnes de publicité antérieures.

Après un travail de pré-processing classique, trois algorithmes classiques sont utilisés pour estimer $\mathbf{p}(\mathbf{x})$: XGBoost [1], Machine à factoriser (FM) [7] et Deep Factorization Machine (DFM) [2]. On peut noter que le premier algorithme se relie assez directement aux forêts aléatoires alors les deux autres sont des modèles linéaires avec des interactions d'ordre deux pour FM et des interactions d'ordre supérieur estimés avec un réseau de neurones pour DFM.

2.2 Optimisation du bid

La stratégie de bid linéaire est la technique la plus répandue dans l'industrie à l'heure actuelle. Elle consiste à ajuster un coefficient multiplicateur λ à la probabilité de clic ($p(x)$) :

$$b_t = \lambda \times \frac{p(x)}{\tau} \tag{1}$$

λ est estimé sur les campagnes passées. Largement utilisée dans l'industrie, cette stratégie d'enchères est utilisée ici comme référence. Elle nécessite cependant une intervention humaine d'ajustement du paramètre λ au jour le jour afin de suivre les évolutions d'un marché hautement dynamique. L'apprentissage par renforcement est une alternative qui exploite la modélisation MDP et permet l'ajustement automatique de la stratégie aux dynamiques de l'environnement.

Au vu des résultats préliminaires, nous nous focalisons sur l'algorithme Actor-Critic (AC) [4, 5]. C'est un algorithme composé de deux réseaux de neurones : un Acteur chargé de choisir une action $\mathbf{a}_t$ au temps t conditionnellement à l'état $\mathbf{s}_t$, et un Critique chargé d'évaluer l'action choisie par l'acteur. Les actions de l'acteur sont mise à jour au fur et à mesure des retours du critique et $\mathbf{a}_t$ converge alors vers la meilleure action en terme d'espérance de gains.

Les données de chaque campagne sont découpées en segments de $T = 1000$ enchères appelés épisodes. Les modèles sont entraînés à la fin de chaque épisode. Quant à la formulation des états, nous utilisons le nombre d'épisodes restants, ici calculé sur le jeu de données mais pouvant être estimé facilement sur les données historiques d'un annonceur. Les états contiennent aussi le rythme de consommation du budget entre t et $t+1$ ainsi que le pourcentage de consommation du budget total alloué à l'épisode en cours.

Dans notre cas, l'action consiste à enchérir plus ou moins que le b_t du bid linéaire. Nous avons choisi l'ensemble des actions $\mathbf{a_t} \in A = \{-0.08, -0.03, -0.01, 0, 0.01, 0.03, 0.08\}$. Ceci nous conduit à définir l'enchère de la façon suivante :

$$b_t = \lambda_t \times \frac{p(x)}{\tau} \text{ avec } \lambda_t = \lambda \times (1 + a_t)$$

3 Notre contribution

L'AUC (Area Under the ROC Curve) est l'une des mesure d'évaluation les plus utilisées en RTB. Nous montrerons les biais induits par cette mesure et nos propositions pour y remédier, en particulier nous montrerons comment évaluer conjointement la prédiction du taux de clic et le prix optimal de l'enchère.

Biais de l'AUC appliquée au RTB : Lorsqu'elle est appliquée à la prédiction du taux de clic dans le contexte du RTB, l'AUC présente notamment les biais suivants :

— Elle donne le même poids aux faux positifs et aux faux négatifs. Dans le contexte du RTB, un faux positif revient à miser pour remporter l'enchère sur un affichage n'entraînant pas de clic et constitue donc une perte sèche sur le budget. A l'inverse, un faux négatif nous conduira à ne pas participer à une enchère sur un affichage suivi d'un clic. Cela représente uniquement la perte d'un clic mais pas de budget. Ces deux cas ne devraient donc pas avoir le même poids dans l'évaluation du modèle.
— L'AUC reflète la performance globale du modèle de prédiction pour tous les seuils de séparation : les régions extrêmes sont donc aussi prises en compte là où, dans le contexte du RTB, ces régions ne seront pas utilisées.
— L'AUC ne prends en compte que l'ordre des prédictions et pas les probabilités en elles-mêmes. Cela n'est pas adapté à la stratégie d'enchère puisque elle est basée sur la probabilité prédite par le modèle.

Nous montrerons comment utiliser l'algorithme Actor-Critic pour l'estimation du coefficient multiplicateur λ_t.

Nous présentons dans le tableau 1 l'efficacité de cette approche comparée à la baseline de l'enchère linéaire largement utilisée dans l'industrie. Ces résultats sont obtenus en appliquant les algorithmes sur le jeu de données publique IpinYou [8]. Ce jeu de données est le plus utilisé et étudié et est organisé neuf campagnes d'enchères, chacune correspondant à un annonceur. Dans la version du jeu de données que nous utilisons [1], chaque campagne comporte entre 156 063 et 350 000 d'historique d'enchères avec un taux moyen de clic entre 0.028% (campagne 2261) et 0.113% (campagne 3358) souvent reliés au type de produits proposés. Le support de ces campagnes est variable et par exemple une campagne sur téléphone mobile génère plus de clics accidentels et donc complique l'estimation de $p(x)$, la probabilité de clic. Dans le tableau 1, **Lin** et **AC** correspondent respectivement à la référence linéaire et à notre version de l'Actor-Critic.

Nous incluons dans ce tableau, les performances d'un algorithme naïf, **Constant**, consistant à enchérir de manière constante $b_t = \lambda$. Ces résultats nous permettent de mettre en évidence les campagnes pour lesquelles l'estimation de la probabilité de clic n'a pas été efficace, comme par exemple sur la campagne 2997 où cette stratégie naïve produit de meilleurs résultats que la stratégie linéaire ou l'algorithme AC.

Le budget disponible pour chaque campagne est fixé à une fraction B du budget nécessaire pour remporter toutes les enchères. Nous étudierons les performances sous différentes contraintes de budget comme il en est coutume dans la communauté. Par souci de simplicité, nous ne présentons ici que les résultats pour $B = 1/32$.

Toutes nos expériences sont conduites en mode asynchrone (*offline*). Cela soulève des questions d'adaptabilité et d'application dans le monde réel puisque le processus d'enchère en temps réel ne doit pas dépasser les 100 millisecondes afin de ne pas nuire à l'expérience de navigation des utilisateurs en allongeant le temps d'affichage des sites. Cette contrainte

1. https ://github.com/wnzhang/make-ipinyou-data

devra donc être prise en compte lors du déploiement de nos algorithmes et s'appuyer sur des méthodes d'apprentissage asynchrone comme le fait l'algorithme A3C (Asynchronous Advantage Actor-Critic) [6].

TABLE 1 – AUC sur la prédiction du CTR et nombre de clics obtenus

camp ID	nb d'enchères	nb de clics potentiels	AUC	Constant	Lin	AC
1458	350 000	304	0.9768	9	247	**261**
2259	350 000	106	0.6877	5	**9**	**9**
2261	343 862	97	0.6216	**10**	8	8
2821	350 000	206	0.6194	16	16	**21**
2997	156 063	530	0.6044	**72**	71	71
3358	300 928	260	0.9758	7	188	**201**
3386	350 000	280	0.7776	1	38	**39**
3427	350 000	230	0.9787	4	145	**156**
3476	350 000	196	0.9579	2	110	**111**

L'importance de la formulation des états et les implications en terme de convergence des algorithmes seront discutées. Nous discuterons aussi de l'élaboration d'une fonction de récompense adaptée à l'application de l'apprentissage par renforcement aux enchères en temps réel.

Références

[1] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.

[2] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm : A factorization-machine based neural network for CTR prediction. *CoRR*, abs/1703.04247, 2017.

[3] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

[4] Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1008–1014. The MIT Press, 1999.

[5] Junwei Lu, Chaoqi Yang, Xiaofeng Gao, Liubin Wang, Changcheng Li, and Guihai Chen. Reinforcement learning with sequential information clustering in real-time bidding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1633–1641, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1928–1937. JMLR.org, 2016.

[7] Steffen Rendle. Factorization machines. In Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu, editors, *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 995–1000. IEEE Computer Society, 2010.

[8] Weinan Zhang, Shuai Yuan, and Jun Wang. Real-time bidding benchmarking with ipinyou dataset. *CoRR*, abs/1407.7073, 2014.

Comparaison des sondages indirects simple et double. Application à l'estimation du trafic postal en France.

Estelle Medous[1,2], Camelia Goga[3], Anne Ruiz-Gazen[1], Jean-François Beaumont[4], Alain Dessertaine[2] et Pauline Puech[2].

[1] *Toulouse School of Economics, Université Toulouse 1 Capitole*
1, Esplanade de l'Université, 31000 Toulouse
E-mail : estelle.medous@laposte.fr, anne.ruiz-gazen@tse-fr.eu
[2] *La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex.*
Email : alain.dessertaine@laposte.fr, pauline.puech@laposte.fr
[3] *Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté*
Email : camelia.goga@univ-fcomte.fr
[4] *Statistique Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada*
Email : jean-francois.beaumont@canada.ca

Résumé. Dans les enquêtes probabilistes, lorsqu'il n'y a pas de base de sondage pour la population cible, une solution consiste à trouver une base de sondage liée à la population cible et à utiliser un échantillonnage indirect. Les poids d'échantillonnage peuvent être déterminés à l'aide de la méthode généralisée du partage des poids (MGPP). Toutefois, cette méthode ne peut pas être appliquée lorsque certains des liens entre la base de sondage et l'échantillon de la population cible sont manquants ou difficiles à récupérer de manière exhaustive. Une solution pour éviter ce problème est de considérer une population intermédiaire liée à la fois à la base de sondage et à la population cible et d'utiliser un double échantillonnage indirect. La MGPP peut alors être utilisée deux fois, d'abord entre la population de base et la population intermédiaire, puis entre la population intermédiaire et la population cible. Comme l'illustre l'enquête française sur le trafic postal, ce double échantillonnage indirect peut détériorer la précision des estimateurs dans certaines situations. Mathématiquement, il est possible de mettre en évidence l'ampleur de la perte de précision dans des situations pratiques proches du contexte de La Poste. Les résultats sont illustrés par des simulations Monte-Carlo et par une application à l'estimation du trafic postal français.

Mots-clés. Enquêtes, Estimation de variance, Méthode généralisée du partage des poids, Plan de sondage complexe, Population finie.

Abstract. In probabilistic surveys, when there is no sampling frame for the target population, a solution is to find a frame population linked in some way to the target population and use indirect sampling. The sampling weights can be determined using the generalized weight share method (GWSM). However, this method cannot be applied when some of the links between the frame population and the sample in the target population

are missing or difficult to retrieve exhaustively. A solution to avoid this issue is to consider an intermediate population linked in some way to both the frame and target populations and use a double indirect sampling. Then the GWSM can be used twice, first between the frame and intermediate populations and then between the intermediate and target populations. As illustrated with the French postal traffic survey, this double indirect sampling appears to be deteriorating the precision of estimators in some situations. Using mathematical derivations, it is possible to highlight the magnitude of the loss of precision in practical situations similar to the French postal context. Results are illustrated through Monte Carlo simulations and with an application to the French postal traffic estimation.

Keywords. Complex sampling design, Finite population, Generalized Weight Share Method, Surveys, Variance estimation.

1 Introduction

En France, l'estimation du trafic postal mensuel par "La Poste" est basé sur un tirage d'échantillon probabiliste. Jusqu'au début des années 2010, les échantillons étaient tirés directement dans la population des tournées de facteurs, qui constitue la population cible ou d'intérêt. Récemment, l'organisation des tournées a évolué de telle façon que cette population n'est plus stable dans le temps. Il n'est plus possible d'échantillonner directement les tournées, et le plan de sondage a été modifié en un tirage dans la population des adresses, qui constitue la base de sondage. Chaque tournée de facteur étant constituée d'adresses, il est possible de relier la population cible à la base de sondage et d'utiliser un plan de sondage indirect pour récupérer un échantillon de tournées.

L'échantillonnage indirect a été étudié de manière intensive dans la littérature (voir par exemple Deville & Lavallée, 2006 et Lavallée, 2007). La méthode d'estimation privilégiée dans ce contexte est la méthode dite méthode généralisée de partage des poids (MGPP). Elle consiste à utiliser les liens qui existent entre la base de sondage et la population cible pour exprimer un total d'intérêt sur la population cible comme un total sur la base de sondage. Les méthodes d'estimation classiques comme l'estimateur d'Horvitz-Thompson peuvent alors être utilisées. La MGPP est une méthode simple mais elle nécessite que les liens entre la base de sondage et la population cible soient connus. Pour l'exemple de La Poste, il s'agit de connaître toutes les adresses dont le courrier est délivré par un facteur lors d'une tournée échantillonnée. On a une moyenne d'environ 500 adresses par tournée et il n'est pas possible de collecter toute l'information avant le départ du facteur. Pour contourner le problème, La Poste a mis en place un sondage doublement indirect en utilisant les casiers de tri du courrier comme une population intermédiaire entre la population des adresses et celle des tournées. Pour ce plan de sondage, il suffit de connaître les casiers des tournées échantillonnées (50 en moyenne) et les adresses du casier associé à l'adresse échantillonnée (10 en moyenne). Avec 60 éléments d'information à collecter en moyenne par tournée échantillonnée pour ce sondage indirect double, au lieu

de 500 pour le sondage indirect simple, il devient possible de mettre en œuvre la méthode d'échantillonnage et ainsi maîtriser les biais d'estimation. Toutefois, La Poste a observé une détérioration importante de la précision des estimateurs de trafic postal après avoir mis en place ce double sondage indirect.

L'objectif de ce travail est d'évaluer la détérioration de la précision en comparant les variances des estimateurs MGPP de totaux entre un sondage indirect simple et double. Ces calculs mathématiques aboutissent à une évaluation précise de la différence de variances dans un contexte proche de celui de La Poste, et permettent d'expliquer la perte de précision observée en pratique. Des simulations de type Monte Carlo valident les résultats et permettent de distinguer des situations où la perte de précision liée à l'utilisation d'un sondage indirect double est faible, voire nulle, de situations où la perte peut-être très forte. Au delà de la compréhension de la perte de précision pour l'estimation du trafic postal, les résultats obtenus permettent de donner des recommandations pour une mise en œuvre efficace d'un plan de sondage indirect double.

Dans la section 2, nous rappelons les notations et les définitions des estimateurs MGPP pour un sondage indirect simple puis double. Dans la section 3, nous donnons les principaux résultats concernant l'étude de la différence de variances des estimateurs Horvitz-Thompson entre sondage indirect simple et double.

2 Sondage indirect

2.1 Sondage indirect simple

Soit y la variable d'intérêt et y_k la valeur de y pour l'individu k dans la population cible U_T. L'objectif est d'estimer le total $t_y = \sum_{k \in U_T} y_k$ de la variable y sur U_T. On suppose que la liste exhaustive des unités de U_T n'est pas disponible mais qu'il existe une base de sondage U_F reliée à U_T de telle façon que chaque unité de U_T soit liée à au moins une unité de U_F. Dans ce cas, l'échantillonnage indirect, tel que détaillé par Deville et Lavallée (2006), permet de sélectionner un échantillon s_F de U_F par un plan de sondage classique, noté p, et d'utiliser des méthodes standards d'estimation de paramètres sur U_T. Dans l'exemple de La Poste, la population cible est constituée des tournées de facteurs, la base de sondages est composée d'adresses et chaque tournée contient au moins une adresse.

Dans un plan indirect, la base de sondage U_F et la population cible U_T peuvent être reliées de diverses manières (voir Deville et Lavallée, 2006, pour plus de détails). Dans le cas de La Poste, pour un jour donné, une adresse n'est délivrée que par une seule tournée (hors organisations dédiées à la distribution des colis) et on parle de liens de type "tous pour un", puisqu'une unité de U_F n'est liée qu'à une seule unité de U_T mais qu'une unité de U_T peut être reliée à plusieurs unités de U_F. Dans la suite de cet article, nous considérons uniquement ce type de liens. A chaque paire (i, k) de $U_F \times U_T$, est associé un indicateur pondéré de lien (ou plus simplement poids de lien), noté θ_{ik}. On a $\theta_{ik} = 0$,

si les unités i et k ne sont pas liées, et un poids strictement positif $\theta_{ik} > 0$ sinon. Pour pouvoir exprimer un total sur la population U_T comme un total pondéré sur U_F, il est nécessaire de normaliser les θ_{ik}. Dans la suite, on note $\tilde{\theta}_{ik} = \theta_{ik} / \sum_{i' \in U_F} \theta_{i'k}$ les poids de liens normalisés et on a $\sum_{i \in U_F} \tilde{\theta}_{ik} = 1$ pour tout k de U_T. On peut alors écrire le total t_y de y sur U_T, comme un total sur U_F pour une variable artificielle $\tilde{y}_i = \sum_{k \in U_T} \tilde{\theta}_{ik} y_k$:

$$t_y = \sum_{k \in U_T} y_k = \sum_{k \in U_T} \left(\sum_{i \in U_F} \tilde{\theta}_{ik} \right) y_k = \sum_{i \in U_F} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k = \sum_{i \in U_F} \tilde{y}_i.$$

Pour estimer t_y, on tire un échantillon s_F dans U_F avec le plan de sondage p et on note $\pi_i = p(i \in s_F)$, $i \in U_F$, les probabilités d'inclusion d'ordre un, supposées strictement positives pour tout $i \in U_F$. L'estimateur Horvitz-Thompson de t_y est donné par :

$$\hat{t}_{y1} = \sum_{i \in s_F} \frac{\tilde{y}_i}{\pi_i} = \sum_{i \in s_F} \frac{1}{\pi_i} \left(\sum_{k \in U_T} \tilde{\theta}_{ik} y_k \right) = \sum_{k \in U_T} \left(\sum_{i \in s_F} \frac{\tilde{\theta}_{ik}}{\pi_i} \right) y_k.$$

Cet estimateur est appelé estimateur MGPP et il est sans biais pour t_y si et seulement si les poids de liens sont normalisés.

Un choix simple pour les poids de liens est $\theta_{ik} = 1$ si i et k sont liés et 0 sinon. Les poids normalisés associés sont donnés par $\tilde{\theta}_{ik} = 1/N_F^k$ où N_F^k est le nombre d'unités de U_F reliées à l'unité k dans U_T. Deville et Lavallée (2006) montrent que pour des plans de sondage tels que le plan de Bernoulli ou le plan aléatoire simple sans remise, ces poids de liens vérifient une propriété d'optimalité. Parmi tous les poids de liens normalisés possibles, ils minimisent la variance de l'estimateur MGPP pour certaines variables d'intérêt (voir la propriété d'optimalité faible dans Deville et Lavallée, 2006).

2.2 Sondage indirect double

Il est possible que certains liens entre la base de sondage et la population cible, nécessaires à l'estimation des paramètres d'intérêt, soient inconnus car trop coûteux à récupérer en pratique, comme dans le cas de La Poste. Dans ce cas, une population intermédiaire U_M permettant de relier U_F à U_T peut être utilisée avec un double sondage indirect et une double mise en œuvre de la MGPP qui permet de diminuer le nombre de liens nécessaires. Comme expliqué précédemment, dans le cas de La Poste, le nombre d'adresses par tournée étant trop important pour être collecté avant le départ des facteurs, la population des casiers de tri du courrier est utilisée comme population intermédiaire entre la population d'adresses et celle des tournées de facteur.

Le sondage indirect simple se généralise sans difficulté au cas du sondage indirect double mais demande d'introduire des notations supplémentaires (voir aussi Deville et Lavallée, 2006, et la propriété de transitivité). Le poids de lien normalisé entre $i \in U_F$ et

$j \in U_M$ (resp. $j \in U_M$ et $k \in U_T$) est noté $\tilde{\theta}_{ij}$ (resp. $\tilde{\theta}_{jk}$). On peut alors écrire le total t_y de y sur U_T comme un total sur U_F pour la variable artificielle $\tilde{\tilde{y}}_i = \sum_{j \in U_M} \tilde{\theta}_{ij} \sum_{k \in U_T} \tilde{\theta}_{jk} y_k$:

$$t_y = \sum_{k \in U_T} y_k = \sum_{k \in U_T} \left(\sum_{j \in U_M} \tilde{\theta}_{jk} \right) \left(\sum_{i \in U_F} \tilde{\theta}_{ij} \right) y_k = \sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij} \sum_{k \in U_T} \tilde{\theta}_{jk} y_k = \sum_{i \in U_F} \tilde{\tilde{y}}_i.$$

On peut aussi en déduire l'estimateur MGPP de t_y :

$$\hat{t}_{y2} = \sum_{i \in s_F} \frac{\tilde{\tilde{y}}_i}{\pi_i} = \sum_{i \in s_F} \frac{1}{\pi_i} \left(\sum_{j \in U_M} \tilde{\theta}_{ij} \sum_{k \in U_T} \tilde{\theta}_{jk} y_k \right) = \sum_{k \in U_T} \left(\sum_{i \in s_F} \frac{1}{\pi_i} \sum_{j \in U_M} \tilde{\theta}_{ij} \tilde{\theta}_{jk} \right) y_k.$$

3 Comparaison des sondages indirects simple et double

Pour comparer les deux plans indirects introduits précédemment, on fixe les populations U_F et U_T ainsi que les paires $(i,k) \in U_F \times U_T$ qui sont liées. On rappelle qu'on ne considère que des liens de type "tous pour un" entre U_F et U_T. On considère les poids de liens normalisés $\tilde{\theta}_{ik}$. Pour le plan indirect double, les poids de liens sont normalisés mais peuvent être quelconques tant qu'ils conduisent aux mêmes paires $(i,k) \in U_F \times U_T$ liées que pour le sondage indirect simple.

Les variances de $\hat{t}_{y1}$ and $\hat{t}_{y2}$ se déduisent facilement des formules de variance de l'estimateur Horvitz-Thompson. Soient $\pi_{ii'} = p(i, i' \in s_F)$, les probabilités d'inclusion d'ordre deux. Les variances de $\hat{t}_{y1}$ et $\hat{t}_{y2}$ sont données par :

$$\mathrm{Var}(\hat{t}_{y1}) = \sum_{i \in U_F} \sum_{i' \in U_F} \frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}} \sum_{k \in U_T} \tilde{\theta}_{ik} y_k \sum_{k' \in U_T} \tilde{\theta}_{i'k'} y_{k'} = \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}})$$

où $\hat{t}_{\tilde{\theta}_k} = \sum_{i \in s_F} \tilde{\theta}_{ik} / \pi_i$ est l'estimateur Horvitz-Thompson du total des poids de liens entre U_F et U_T, pour l'individu k dans U_T, dans le cas du sondage indirect simple et

$$\begin{aligned} \mathrm{Var}(\hat{t}_{y2}) &= \sum_{i \in U_F} \sum_{i' \in U_F} \frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}} \sum_{k \in U_T} \sum_{j \in U_M} \tilde{\theta}_{ij} \tilde{\theta}_{jk} y_k \sum_{k' \in U_T} \sum_{j' \in U_M} \tilde{\theta}_{i'j'} \tilde{\theta}_{j'k'} y_{k'} \\ &= \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \mathrm{Cov}(\hat{t}_{\tilde{\tilde{\theta}}_k}, \hat{t}_{\tilde{\tilde{\theta}}_{k'}}) \end{aligned}$$

où $\hat{t}_{\tilde{\tilde{\theta}}_k} = \sum_{i \in s_F} (\sum_{j \in U_M} \tilde{\theta}_{ij} \tilde{\theta}_{jk}) / \pi_i$ est l'estimateur Horvitz-Thompson du total des poids de liens entre U_F et U_T, pour l'individu k dans U_T, dans le cas du sondage indirect double. Pour comparer la précision du sondage indirect simple avec le double, nous calculons la différence des variances :

$$\mathrm{Var}(\hat{t}_{y2}) - \mathrm{Var}(\hat{t}_{y1}) = \sum_{k \in U_T} \sum_{k' \in U_T} y_k y_{k'} \left(\mathrm{Cov}(\hat{t}_{\tilde{\tilde{\theta}}_k}, \hat{t}_{\tilde{\tilde{\theta}}_{k'}}) - \mathrm{Cov}(\hat{t}_{\tilde{\theta}_k}, \hat{t}_{\tilde{\theta}_{k'}}) \right).$$

La proposition suivante permet de simplifier l'expression de la différence de variances de $\hat{t}_{y2}$ et $\hat{t}_{y1}$ dans le cas d'un plan de Bernoulli ou aléatoire simple sans remise.

Proposition 1 *Si p est un plan de Bernoulli ou un plan aléatoire simple sans remise, si les liens entre U_F et U_T sont de type 'tous pour un" et si on considère, pour tout k de U_T, les poids de liens $\tilde{\theta}_{ik} = 1/N_F^k$ où N_F^k est le nombre d'unités de U_F reliées à k, nous avons :*

$$Var\ (\hat{t}_{y2}) - Var\ (\hat{t}_{y1}) = \sum_{k \in U_T} y_k^2 Var\ (\hat{t}_{\tilde{\tilde{\theta}}_k} - \hat{t}_{\tilde{\theta}_k})$$

On déduit de la proposition précédente que pour les plans usuels de Bernoulli et aléatoire simple sans remise, dans le cas de liens de type "tous pour un", le plan indirect double conduit toujours à des estimateurs MGPP moins précis que l'estimateur obtenu par sondage indirect simple avec des poids de liens optimaux. De plus, on voit que cette différence de variances est liée à la variabilité de la différence des estimateurs des poids de liens. Selon la façon dont la base de sondage est liée à la population intermédiaire et la population intermédiaire à la population cible, on peut obtenir des différences de variances faibles ou fortes. Différents scénarios seront envisagés et des simulations Monte Carlo permettront d'illustrer les situations où le plan indirect double conduit à une perte de précision importante comparé au plan indirect simple. Ces résultats peuvent se généraliser au plan de Poisson pour des poids de liens optimaux ainsi qu'aux plans stratifiés aléatoires simples sans remise sous certaines conditions. L'application de ces résultats dans un contexte proche de celui de La Poste sera aussi présentée.

Bibliographie

[1] Deville, J.-C. et Lavallée, P. (2006). Indirect sampling : the foundations of the generalized weight share method, *Survey methodology*, 32(2), 165-176.

[2] Lavallée, P. (2007). *Indirect sampling*, Springer-Verlag New York.

Clustering parcimonieux pour extrêmes multivariés

Nicolas MEYER [1] & Olivier WINTENBERGER [2]

[1] *Sorbonne Université, LPSM, France*
meyer@math.ku.dk
[2] *Sorbonne Université, LPSM, France*
olivier.wintenberger@upmc.fr

Résumé. Identifier les directions dans lesquelles des événements exceptionnels apparaissent est un des problèmes majeurs de la théorie multivariée des valeurs extrêmes. D'un point de vue théorique, la majeure partie de l'information concernant de tels événements est contenue dans la mesure spectrale, qui apparaît comme la limite de la composante angulaire de vecteurs aléatoires à variation régulière. Estimer cette mesure s'avère être un point délicat, notamment en grande dimension. Dans cette présentation, nous introduisons une méthode de réduction de la dimension basée sur la projection euclidienne sur le simplexe. Cette projection a été étudiée dans le cadre des valeurs extrêmes par Meyer & Wintenberger (2021) qui ont établi plusieurs résultats théoriques. La présentation s'attachera à exposer une approche statistique basée sur de la sélection de modèle qui permet d'identifier les groupes de coordonnées susceptibles d'être extrêmes simultanément. Cette approche donne lieu à un algorithme appelé *MUSCLE* pour *MUltivariate Sparse Clustering for Extremes*.

Mots-clés. Extrêmes multivariés, measure spectrale, projection sur le simplexe.

Abstract. Identifying directions where exceptional events occur is one of the major problems of multivariate extreme value theory. From a theoretical point of view most of the information concerning such events is contained in the spectral measure which appears as the limit of the angular component of regularly varying random vectors. Estimating this measure is a delicate point especially in large dimensions. In this presentation we introduce a dimension reduction method based on the Euclidean projection onto the simplex. This projection has been studied in the context of extreme values by Meyer & Wintenberger (2021) who established several theoretical results. The presentation will focus on a statistical approach that uses model selection to identify groups of coordinates that are likely to be extreme simultaneously. This approach gives rise to an algorithme called *MUSCLE* for *MUltivariate Sparse Clustering for Extremes*.

Keywords. Multivariate extremes, projection onto the simplex, spectral measure.

1 Valeurs extrêmes et variation régulière

1.1 Variation régulière

Étudier les valeurs extrêmes générées par un vecteur aléatoire $\mathbf{X} \in \mathbb{R}_+^d$, $d \geq 2$, revient à étudier le comportement de la queue de distribution de $\mathbf{X}$. Dans ce contexte, il est courant de supposer que le vecteur $\mathbf{X}$ est à variation régulière : il existe un vecteur aléatoire $\mathbf{\Theta}$ sur la sphère unité telle que

$$\mathbb{P}\big((|\mathbf{X}|/t, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > t\big) \xrightarrow{d} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,, \tag{1}$$

cf Resnick (2007). Dans ce cas, $|\cdot|$ désigne n'importe quelle norme sur $\mathbb{R}^d$. Le vecteur limite $\mathbf{\Theta}$ est alors appelé vecteur spectral tandis que sa loi est appelée mesure spectrale. La convergence (1) permet de séparer l'étude de la composante radiale des extrêmes $|\mathbf{X}|/t$ de celle de la composante angulaire $\mathbf{X}/|\mathbf{X}|$. Cette dernière concentre l'information sur la localisation et la dépendance des valeurs extrêmes. L'étude de la mesure spectrale est donc un point central de la théorie des extrêmes multivariés.

Il est souvent intéressant (voir par exemple Goix et al. (2017)) d'étudier le comportement du vecteur spectral sur les sous-ensembles C_β de la sphère définis par

$$C_\beta = \{\mathbf{x} \in \mathbb{R}_+^d : |\mathbf{x}| = 1,\, x_j > 0 \text{ pour } j \in \beta,\, x_j = 0 \text{ pour } j \notin \beta\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. Des groupes de directions β sont appelés *clusters*. En effet, la mesure spectrale met de la masse sur un tel ensemble si des événements extrêmes apparaissent conjointement dans la direction β. On est ainsi ramené à l'estimation des probabilités $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$. Cependant, l'estimation de ces quantités se révèle délicate pour essentiellement deux raisons. Tout d'abord, le nombre de probabilités à estimer croît exponentiellement en la dimension. Par ailleurs, si $\beta \neq \{1, \ldots, d\}$ vérifie $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, alors la mesure spectrale charge la frontière de C_β (qui est le sous-ensemble C_β lui-même) et donc la convergence (1) ne s'applique pas.

L'idée proposée par Meyer et Wintenberger (2021) pour contourner ce problème est de remplacer le vecteur unitaire $\mathbf{X}/|\mathbf{X}|$ de (1) par un autre projeté qui permet de mieux tenir compte de la masse mise par la mesure spectrale sur les sous-ensembles C_β. Cette modification de la convergence (1) donne alors naissance à la notion de variation régulière parcimonieuse.

1.2 Variation régulière parcimonieuse

Introduite principalement par Duchi et al. (2008), la projection euclidienne sur le simplexe a connu un usage divers et varié, notamment en théorie de l'apprentissage.

Dans la suite, $|\cdot|$ désigne la norme ℓ^1 et $\mathbb{S}_+^{d-1}$ désigne le simplexe de $\mathbb{R}^d$. Si $\mathbf{v} \in \mathbb{R}_+^d$, alors le vecteur projeté $\pi(\mathbf{v})$ est l'unique vecteur $\mathbf{w}$ de $\mathbb{S}_+^{d-1}$ qui minimise la quantité

$|\mathbf{w} - \mathbf{v}|_2$, où $|\cdot|_2$ désigne la norme ℓ^2. On note alors π la projection euclidienne sur le simplexe $\mathbb{S}_+^{d-1}$. Cette manière de projeter permet de rendre les vecteurs parcimonieux, c'est-à-dire avec plusieurs coordonnées nulles. Elle permet ainsi de mieux rendre compte du comportement des extrêmes sur les ensembles C_β.

Définition 1 (Variation régulière parcimonieuse)**.** *Un vecteur* $\mathbf{X}$ *à valeurs dans* $\mathbb{R}_+^d$ *est dit à variation régulière parcimonieuse s'il existe un vecteur aléatoire* $\mathbf{Z}$ *défini sur le simplexe et une variable aléatoire positive* Y *tels que*

$$\mathbb{P}\big((|\mathbf{X}|/t, \pi(\mathbf{X}/t)) \in \cdot \mid |\mathbf{X}| > t\big) \xrightarrow{d} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{2}$$

Le vecteur limite $\mathbf{Z}$ doit être vu comme la limite angulaire obtenue après avoir remplacé $\mathbf{X}/|\mathbf{X}|$ par $\pi(\mathbf{X}/t)$ dans l'Equation (1). Par continuité de la projection, la notion de variation régulière standard (Equation (1)) implique celle de variation régulière parcimonieuse. Meyer et Wintenberger (2021) ont prouvé que sous des hypothèses assez faibles, les deux notions sont en fait équivalentes.

L'intérêt principal de la Définition 1 est de pouvoir approcher le comportement des extrêmes de $\mathbf{X}$ sur les ensembles C_β. En effet, en reprenant les notations précédentes, on a la convergence suivante pour tout $\beta \subset \{1, \dots, d\}$:

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \quad t \to \infty\,. \tag{3}$$

L'objectif est donc d'estimer le support de la distribution de $\mathbf{Z}$ via l'estimation des probabilités $\mathbb{P}(\mathbf{Z} \in C_\beta)$, pour $\beta \subset \{1, \dots, d\}$, le but étant de détecter lesquelles de ces probabilités sont positives. Autrement dit, il s'agit d'identifier l'ensemble

$$\mathcal{S}^*(\mathbf{Z}) := \{\beta \subset \{1, \dots, d\} : \mathbb{P}(\mathbf{Z} \in C_\beta) > 0\}\,.$$

Cet ensemble $\mathcal{S}^*(\mathbf{Z})$ rassemble tous les clusters de directions β sur lesquelles le vecteur angulaire $\mathbf{Z}$ met de la masse. On note s^* son cardinal. L'objectif est alors de proposer une approche statistique pour décider quels clusters β appartiennent à $\mathcal{S}^*(\mathbf{Z})$.

2 Estimation

On considère désormais une suite de vecteurs aléatoires indépendants et identiquement distribués à variation régulière $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \dots$ de vecteur spectral $\mathbf{\Theta}$. On considère également une variable aléatoire Y de loi de Pareto de paramètre $\alpha > 0$, indépendante de $\mathbf{\Theta}$. Enfin, on pose $\mathbf{Z} = \pi(Y\mathbf{\Theta})$.

Le cadre classique en statistique des valeurs extrêmes est de considérer une suite positive $(u_n)_{n\in\mathbb{N}}$ telle que $u_n \to \infty$. Cette suite joue le rôle du seuil t dans les Equations (1) et (2). Cela signifie que pour $n \in \mathbb{N}$, la quantité u_n doit être vue comme le seuil au-dessus duquel les données $\mathbf{X}_1, \cdots, \mathbf{X}_n$ sont considérées comme des valeurs extrêmes.

Il est également usuel de définir un niveau $k = k_n = n\mathbb{P}(|\mathbf{X}| > u_n)$ et de supposer que $k_n \to \infty$ quand $n \to \infty$. Il est à noter que l'hypothèse $u_n \to \infty$ implique que $k_n/n = \mathbb{P}(|\mathbf{X}| > u_n) \to 0$. Ainsi, k_n tend vers l'infini à une vitesse plus lente que n. Un estimateur naturel non biaisé pour k_n est $\hat{k} = \hat{k}_n = \sum_{j=1}^n \mathbb{1}_{|\mathbf{X}_j|>u_n}$ qui correspond au nombre de dépassements au-dessus du seuil u_n, c'est-à-dire au nombre de valeurs extrêmes.

Notre objectif est d'estimer les probabilités $p^*(\beta) := \mathbb{P}(\mathbf{Z} \in C_\beta)$ pour $\beta \subset \{1, \ldots, d\}$. Ces probabilités apparaissent comme les limites des probabilités pré-asymptotiques $p_n(\beta) := \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$ (voir Equation (3)). Le problème principal est alors de décider si $p(\beta)$ est positif ou nul. On définit pour cela l'estimateur

$$T_n(\beta) := \sum_{j=1}^{n} \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. L'idée est de sélectionner parmi les valeurs extrêmes celles qui sont projetées dans l'ensemble C_β. Ces estimateurs vérifient les résultats asymptotiques suivants.

Théorème 1. *On reprend les notations précédentes.*

1. *(Consistance). Le vecteur $k_n^{-1}(T_n(\beta))_{\beta\subset\{1,\ldots,d\}}$ converge en probabilité vers $\mathbf{p}^* := (p(\beta))_{\beta\subset\{1,\ldots,d\}}$.*
2. *(Normalité asymptotique). On a la convergence en loi suivante :*

$$\sqrt{k_n}\,\mathrm{Diag}\left(\mathbf{p}_{\mathcal{S}^*(\mathbf{Z})}\right)^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{S}^*(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{S}^*(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,.$$

où $\mathbf{T}_{n,\mathcal{S}^*(\mathbf{Z})}$ (resp. $\mathbf{p}_{n,\mathcal{S}^*(\mathbf{Z})}$ et $\mathbf{p}^*_{\mathcal{S}^*(\mathbf{Z})}$) correspond au vecteur de $\mathbb{R}^{s^*}$ dont les composantes sont les $T_n(\beta)$ (resp. $p_n(\beta)$ et $p^*(\beta)$) pour $\beta \in \mathcal{S}^*(\mathbf{Z})$ (on se restreint aux probabilités positives).

3 Sélection de modèle

La répartition des k données extrêmes sur les $2^d - 1$ sous-ensembles $(C_\beta)_{\beta\subset\{1,\ldots,d\}}$ suggère d'utiliser le modèle multinomial $\mathbf{M}_k$ sur R^{2^d-1} de vecteur de probabilités $\mathbf{p}$ défini par

$$\mathbf{p} = (\overbrace{p_1, \ldots, p_s, \underbrace{p, \ldots, p}_{r-s}, 0, \ldots, 0}^{2^d-1 \text{ composantes}})\,,$$

avec $p_1 \geq \ldots \geq p_s, p \in (0,1)$ satisfaisant la contrainte :

$$p_1 + \ldots + p_s + (r-s)\tilde{p} = 1\,.$$

L'idée est de séparer les clusters β en trois catégories. La première correspond aux clusters sur lesquels les extrêmes apparaissent, leur probabilité d'apparition étant alors p_j. La deuxième catégorie correspond aux clusters qui concentrent peu de données extrêmes. Ce phénomène résulte du biais entre l'aspect non-asymptotique de l'étude et le modèle théorique des $p^*(\beta)$. Ce biais est modélisé par une probabilité d'occurence p considérée comme proche de 0. Enfin, la dernière catégorie concerne les clusters β sur lesquels aucune donnée n'est apparue ; on estime alors que ces clusters ne concentrent pas de valeurs extrêmes (d'où une probabilité d'occurrence nulle). L'objectif est alors d'ajuster au mieux le nombre s de faces significatives. Cet ajustement s'effectue par le calcul de la divergence de Kullback-Leibler entre les données et le modèle théorique $\mathbf{M}_k$. L'estimation de cette vraisemblance implique que le modèle qui correspond le mieux aux données est celui qui minimise une log-vraisemblance pénalisée.

La sélection des clusters contenant les extrêmes de $\mathbf{X}$ s'est faite jusqu'à présent pour le choix d'un seuil u_n arbitraire. L'idée est alors d'inclure le choix de ce seuil dans la sélection de modèle. L'approche utilisée consiste à considérer des modèles avec un nombre de valeurs extrêmes différent et de déterminer le modèle le plus approprié. Cette analyse est réalisée en partitionnant les données en un groupe d'extrêmes et un groupe de non-extrêmes, le but de la sélection de modèle étant d'identifier la partition qui correspond le mieux aux données.

Notre approche donne lieu à un algorithme appelé MUSCLE pour *MUltivariate Sparse Clustering for Extremes* qui donne les clusters extrêmes à partir de données $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Cet algorithme ne nécessite aucun hyper-paramètre, contrairement aux méthodes existantes dans la littérature. Ce travail est par ailleurs le premier qui combine l'étude de la dépendance des extrêmes et le choix du seuil.

La fin de notre présentation est consacrée à l'étude de divers exemples sur des données simulées qui illustrent la pertinence de notre approche. Enfin, nous mettons en évidence les clusters extrêmes sur des données financières et environnementales.

Bibliographie

Duchi, J. Shalev-Shwartz, S. Singer, Y. Chandra, T. (2008), *Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions*, ICML.
Goix, N. Sabourin, A. Clémençon, S. (2017), *Sparsity in Multivariate Extremes with Applications to Anomaly Detection*, Journal of Multivariate Analysis.
Meyer, N. et Wintenberger, O. (2021), Sparse regular variation, à paraître dans *Advances in Applied Probability*, arXiv : 1907.00686.
Resnick, S.I. (2007), *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer.

Test de détection de rupture dans un modèle de régression

Zaher Mohdeb

Ecole Nationale Polytechnique de Constantine
et Laboratoire de Mathématiques et Sciences de la Décision
Université frères Mentouri de Constantine, Algérie
E-mail: z.mohdeb@gmail.com

Résumé. On considère un modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixé, notre but est de construire le test de l'hypothèse linéaire contre les alternatives de ruptures de modèle et ce sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative. On établit la normalité asymptotique de la statistique de test sous l'hypothèse nulle ainsi que sous l'hypothèse alternative de ruptures de modèle.

Mots-clés. Hypothèse linéaire, régression non paramétrique, rupture de modèle.

Abstract. We consider a regression model in the case of a homoscedastic error structure and fixed design, our aim is to build the test of the linear hypothesis versus regime switching models without regularity condition, and also under either the null or the alternative hypotheses. We establish the asymptotic normality of the test statistic under the null hypothesis and the alternative one.

Keywords. Linear hypothesis, nonparametric regression, regime switching.

1 Introduction

On considère le modèle de régression suivant

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\,, \quad i = 1, \ldots, n\,, \tag{1}$$

où f est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{i,n}$, $i = 1, \ldots, n$, est un échantillonnage fixé de l'intervalle $[0,1]$. Les erreurs $\varepsilon_{i,n}$ forment un tableau triangulaire de variables aléatoires d'espérance nulle et de variance finie σ^2.
Soient $g_1, \ldots, g_p$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit E_p l'espace vectoriel engendré par $g_1, \ldots, g_p$. On veut tester l'hypothèse:

$$H_0 :\ f \in E_p \quad \text{contre} \quad H_1 : \begin{cases} \exists\, s \in]0,1[\ \text{ tel que } \ f = \phi \mathbb{1}_{[0,s]} + \psi \mathbb{1}_{]s,1]}, \\ \phi \in E_p\,, \ \ \psi \text{ Riemann intégrable et } f \notin E_p\,. \end{cases} \tag{2}$$

La plupart des travaux sur les tests d'hypothèses dans le modèle (1) supposent des conditions de régularité sur f, $g_1, \ldots, g_p$; généralement ces fonctions sont supposées höldériennes. On peut citer, sans être exhaustif, Cox et al (1988), Eubank et Spiegelmann (1990), Eubank et Hart (1992), Azzalini et Bowman (1993), Härdle et Mammen (1993). Les tests basés sur l'estimation sur la L^2-distance entre f et E_p sont étudiés par Dette et Munk ((1998), Munk et Dette (1998), Mohdeb et Mokkadem (2004), avec l'hypothèse que f est höldérienne d'ordre $\gamma > 1/2$.

Dans ce travail, on applique l'approche utilisée dans Mohdeb et Mokkadem (2015) et Lessak et Mohdeb (2015) pour construire le test d'hypthèses (2) dans le modèle (1). On suppose que f, $g_1, \ldots, g_p$ sont Riemann-intégrables; sous cette seule condition sur les fonctions, on établit la normalité asymptotique de la statistique de test qui permet de construire le test (2) et d'avoir la puissance pour des alternatives de ruptures de modèle.

Dans la section 2, on introduit les hypothèses et on présente notre résultat principal.

2 Hypothèses et résultats

On considère le modèle de régression (1) et E_p est l'espace vectoriel engendré par des fonctions fixées $g_1, \ldots, g_p$ définies sur $[0, 1]$ et linéairement indépendantes.

Nos hypothèses sont les suivantes:

- (A1) $\max\limits_{i=2,\ldots,n} \left|(t_{i,n} - t_{i-1,n}) - \dfrac{1}{n}\right| = o\left(\dfrac{1}{n}\right)$;
- (A2) $\forall n$, $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes et $\exists C \in \mathbb{R}^+$ tel que $E(\varepsilon_{i,n}^4) < C$, $\forall i, n$;
- (A3) La fonction f est Riemann-intégrable.
- (A4) Les fonctions $g_1, \ldots, g_p$ sont localement höldériennes d'ordre $\gamma > 1/2$.

Les fonctions que nous considérons, sont aussi dans $L^2(dt)$ muni de son produit scalaire usuel. On pose,

$$\mathcal{D}^2(f) := \min_{v \in E_p} \|f - v\|^2 \tag{3}$$

la distance entre f et le sous-espace E_p, $Y := (Y_{1,n}, \ldots, Y_{n,n})'$, $f_n := \Big(f(t_{1,n}), \ldots, f(t_{n,n})\Big)'$, $g_{k,n} := \Big(g_k(t_{1,n}), \ldots, g_k(t_{n,n})\Big)'$, $k = 1, \ldots, p$, et $G := (g_{1,n}, \ldots, g_{p,n})$.
On note aussi $E_{p,n}$, le sous-espace de $\mathbb{R}^n$ engendré par $\{g_{1,n}, \ldots, g_{p,n}\}$ qui est une discrétisation du sous-espace E_p, $\Pi_n = G(G'G)^{-1}G'$, la matrice de projection sur $E_{p,n}$ et $\Pi_n^{\perp} = I_n - G(G'G)^{-1}G'$, la matrice de projection sur l'espace orthogonal de $E_{p,n}$, où I_n est la matrice identité $n \times n$.

On considère la statistique suivante définie par

$$D_n^2 := \frac{1}{n} Y' \Pi_n^{\perp} Y \,. \tag{4}$$

On vérifie que

$$E(D_n^2) = \widetilde{D}_n^2 + \frac{n-p}{n}\sigma^2 \,, \quad \text{où} \quad \widetilde{D}_n^2 = \frac{1}{n} f_n' \Pi_n^{\perp} f_n \,.$$

On est amené ainsi à considérer $D_n^2 - \frac{n-p}{n}\sigma^2$, mais σ^2 est inconnu. On l'estime à l'aide de l'estimateur suivant, introduit par Gasser, Sroka, et Jennen-Steinmetz (1986)

$$S_\varepsilon^2 = \frac{1}{6(n-2)} \sum_{i=2}^{n-1} \left(Y_{i+1,n} + Y_{i-1,n} - 2Y_{i,n}\right)^2 \,. \tag{5}$$

On obtient ainsi la statistique de test donnée par

$$\widehat{D}_n^2 = D_n^2 - \frac{n-p}{n} S_\varepsilon^2 \,;$$

et on rejette l'hypothèse H_0 : "$f \in E_p$" si $\widehat{D}_n^2 > u_\alpha$, où u_α est un nombre réel positif. Notre résultat principal est le suivant.

Théorème 1 *Si les conditions* $(A1)$, $(A2)$ *et* $(A3)$ *sont satisfaites, alors*

$$\sqrt{n}\left\{\widehat{D}_n^2 - \widetilde{D}_n^2 + B_n(f)\right\} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4 + 4\sigma^2 \mathcal{D}^2(f)\right) \,,$$

où $B_n(f) = \dfrac{1}{6n} \displaystyle\sum_{i=2}^{n-1} \Big(f(t_{i+1,n}) + f(t_{i-1,n}) - 2f(t_{i,n})\Big)^2$.

Nous avons le résultat suivant, comme conséquence du Théorème 1, sous l'hypothèse nulle H_0.

Corollaire 1 *Si les conditions* $(A1)$-$(A4)$ *sont satisfaites, alors sous l'hypothèse nulle* H_0, *on a*

$$\sqrt{n}\,\widehat{D}_n^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4\right) .$$

Ce corollaire donne le niveau asymptotique du test et le Théorème 1 donne la puissance du test pour les hypothèses alternatives de ruptures de modèles. En pratique la variance σ^2 des erreurs est généralement inconnue, on peut considérer un estimateur consistent $\widehat{\sigma}^2$ de σ^2. On rejette l'hypothèse nulle H_0 : "$f \in E_p$", si

$$\frac{\sqrt{n}}{\widehat{\sigma}^2}\,\widehat{D}_n^2 \;>\; z_{1-\alpha} \,,$$

où $z_{1-\alpha}$ est le $(1-\alpha)$quantile d'une loi normale standard.

3 Simulations

Dans nos simulations, on étudie le test de l'hypothèse

$$H_0 : \quad f(t) = t\mathbb{1}_{[0,1]}(t) \quad \text{contre} \quad H_1 : f(t) = t\mathbb{1}_{[0,s]}(t) + \beta t\mathbb{1}_{]s,1]}(t)\,, \quad \beta \neq 1\,,$$

au niveau de signification $\alpha = 0.05$.
On a mené une étude Monte Carlo en simulant le modèle (1), avec $t_{i,n} = \frac{i-1}{n-1}$, $i = 1, \ldots, n$ et une taille d'échantillon $n = 64$ et $\varepsilon_{i,n} \sim \text{i.i.d.}\mathcal{N}(0, \sigma^2)$. La statistique de test est définie par

$$\widehat{D}_n^2 = \frac{1}{n}\sum_{i=1}^{n} |Y_{i,n} - \widehat{a}\, t_{i,n}|^2 - \frac{n-1}{n} S_\varepsilon^2\,, \quad \text{où} \quad \widehat{a} = \frac{\sum_{i=1}^n t_{i,n} Y_{i,n}}{\sum_{i=1}^n t_{i,n}^2}\,.$$

L'hypothèse H_0 est rejetée si

$$\left(\frac{9n}{17}\right)^{1/2} \frac{\widehat{D}_n^2}{\widehat{\sigma}^2} > v_{0.95}\,,$$

où $v_{0.95} = 1.65$ est le quantile d'ordre 0.95 d'une loi normale standard et

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_{i,n} - \widehat{a} t_{i,n})^2\,.$$

Les simulations sont conduites pour différentes valeurs de s, β et σ^2. Les résultats obtenus montrent que, pour les petites valeurs de l'écart-type σ des erreurs, la puissance empirique du test est proche de 1. Les résultats montrent également que pour $\beta = 1$ et $s = 1$, la puissance empirique est proche du niveau de signification $\alpha = 0.05$, car, dans ce cas, l'hypothèse alternative correspond à l'hypothèse nulle H_0.

Bibliographie

[1] Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser. B*, **55**, 549-557.
[2] Cox, D., Koh, G., Wahba, G. and Yandell, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, **16** 113-119.
[3] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 2, 778-800.
[4] Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Stat.*, **20**, 3, 1412-1425.
[5] Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.*, **85**, 410, 387-392.
[6] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika,* **73**, 625-633.

[7] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 4, 1926-1947.
[8] Lessak, R. and Mohdeb, Z. (2015). Testing the linear regression model null hypothesis versus regime switching alternatives. *Afr. Stat.*, **10**, 807-813.
[9] Mohdeb, Z. and Mokkadem, A. (2004). Average squared residuals approach for testing linear hypothesis in nonparametric regression. *J. Nonparametric Stat.*, **16**, 1-2, 3-12.
[10] Mohdeb, Z. and Mokkadem, A. (2015). Testing linear regression models in non regular case. *Comm. Statist. Theory Methods*, **44**, 21, 4476-4490.
[11] Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *Ann. Stat.*, **26**, 6, 2339-2368.

Statistical deconvolution of the free Fokker-Planck equation at fixed time

Mylène Maïda [1] & Tien Dat Nguyen [2] & Thanh Mai Pham Ngoc [3]
& Vincent Rivoirard [4] & Viet Chi Tran [5]

[1] *Université Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, mylene.maida@univ-lille.fr*
[2] *Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, tien-dat.nguyen@math.u-psud.fr*
[3] *Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, thanh.pham_ngoc@math.u-psud.fr*
[4] *CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research Uni., Vincent.Rivoirard@dauphine.fr*
[5] *LAMA, Université Gustave Eiffel, Université Paris Est Creteil, CNRS, chi.tran@univ-eiffel.fr*

Résumé. Nous nous intéressons à la reconstruction de la condition initiale d'une équation aux dérivées partielles non linéaires (EDP), à savoir l'équation de Fokker-Planck, à partir de l'observation d'un mouvement brownien de Dyson à un temps donné $t > 0$. L'équation de Fokker-Planck peut-être obtenue comme la limite d'un système de particules avec répulsion électrostatique correctement renormalisé. La solution de l'équation de Fokker-Planck peut s'écrire comme la convolution libre de la condition initiale et de la distribution de la loi semi-circulaire. Nous proposons un estimateur non-paramétrique de la condition initiale, obtenu en effectuant une déconvolution libre au moyen de fonctions de subordination. Cet estimateur est original car il implique la résolution d'une équation du point fixe et une déconvolution classique par une distribution de Cauchy. En effet, en probabilité libre, l'analogue de la transformée de Fourier est la transformée R, liée à la transformée de Cauchy. La convergence de l'estimateur est prouvée et l'erreur quadratique moyenne intégrée est obtenue, avec des vitesses de convergence similaires à celles connues pour des problèmes d'estimation non paramétrique de densité dans un cadre de déconvolution classique. Enfin, une étude de simulations illustre les bonnes performances de notre estimateur.

Mots-clés. EDP avec condition initiale aléatoire, Déconvolution libre, Problème inverse, Estimation non paramétrique par méthodes à noyau, Transformée de Fourier, MISE, Mouvement brownien de Dyson.

Abstract. We are interested in reconstructing the initial condition of a non-linear partial differential equation (PDE), namely the Fokker-Planck equation, from the observation of a Dyson Brownian motion at a given time $t > 0$. The Fokker-Planck equation describes the evolution of electrostatic repulsive particle systems, and can be seen as the large particle limit of correctly renormalized Dyson Brownian motions. The solution of

the Fokker-Planck equation can be written as the free convolution of the initial condition and the semi-circular distribution. We propose a nonparametric estimator for the initial condition obtained by performing the free deconvolution via the subordination functions method. This statistical estimator is original as it involves the resolution of a fixed point equation, and a classical deconvolution by a Cauchy distribution. This is due to the fact that, in free probability, the analogue of the Fourier transform is the R-transform, related to the Cauchy transform. The convergence of the estimator is proved and the integrated mean square error is computed, providing rates of convergence similar to the ones known for non-parametric deconvolution methods. Finally, a simulation study illustrates the good performances of our estimator.

Keywords. PDE with random initial condition, Free deconvolution, Inverse problem, Non-parametric kernel estimation, Fourier transform, Mean Integrated Error, Dyson Brownian motion.

1 Introduction

We consider the 1-d free Fokker-Planck equation:

$$\frac{\partial}{\partial t}\mu_t = -\frac{\partial}{\partial x}\big[\mu_t\big(H\mu_t\big)\big], \tag{1}$$

where $\big(\mu_t\big)_{t\geq 0}$ is a family of probability measures on $\mathbb{R}$, and initial condition $\mu_0(x) = p_0(x)\mathrm{d}x$, and $\big(H\mu\big)$ denotes the Hilbert transform of a probability measure μ on $\mathbb{R}$, defined for any $x \in \mathbb{R}$ by: $H\mu(x) := \lim_{\varepsilon\to 0}\int_{\mathbb{R}\setminus[x-\varepsilon,x+\varepsilon]} \dfrac{1}{x-y}\mathrm{d}\mu(y)$.
Let $\boxplus$ denote the free convolution between two probability measures on $\mathbb{R}$, which has been first introduced by D. Voiculescu in 1986. Then, the solution μ_t of (1) can be written as:

$$\mu_t = \mu_0 \boxplus \sigma_t, \tag{2}$$

where σ_t is the semi-circular distribution of variance t with density $f_{\sigma_t}(x) = \dfrac{1}{2\pi t}.\sqrt{4t-x^2}$, $x \in [-2\sqrt{t}, 2\sqrt{t}]$.
Observations: we observe a matrix $X^n(t)$ for a given time $t > 0$, t is assumed to be fixed in the sequel, where

$$X^n(t) = X^n(0) + H^n(t),$$

with $X^n(0)$ a diagonal matrix whose entries are the ordered statistic $\lambda_1^n(0) < \cdots < \lambda_n^n(0)$ of a vector $(d_j^n)_{j\in\{1,\dots n\}}$ of n independent and identically distributed (i.i.d.) random variables distributed as μ_0, and $H^n(t)$ a standard Hermitian Brownian motion.
We emphasize that we do not observe directly the initial condition $X^n(0)$. The observation

consists in $X^n(t)$, from which we can compute the eigenvalues $(\lambda_1^n(t), \cdots, \lambda_n^n(t))$ and then the associated empirical measure. It is known that, (see e.g. in [1, Theorem 4.3.2]), the eigenvalues $(\lambda_1^n(t), \cdots, \lambda_n^n(t))$ of $X^n(t)$ solve the following system of stochastic differential equations (SDE):

$$\mathrm{d}\lambda_j^n(t) = \frac{1}{\sqrt{n}}\mathrm{d}\beta_j(t) + \frac{1}{n}\sum_{k\neq j}\frac{\mathrm{d}t}{\lambda_j^n(t) - \lambda_k^n(t)}, \quad 1 \leq j \leq n,$$

where β_j are i.i.d. standard real Brownian motions. Now, we denote by $\mu_t^n = \frac{1}{n}\sum_{j=1}^n \delta_{\lambda_j^n(t)}$, the empirical measure of these eigenvalues at time t, then, under specific conditions, the process $(\mu_t^n)_{t\geq 0}$ converges weakly almost surely as n goes to infinity to the process $(\mu_t)_{t\geq 0}$ solution of (1) with density $(p(t,\cdot))_{t\geq 0}$.
The purpose is to estimate the density p_0 from the data $(\lambda_1^n(t), \cdots, \lambda_n^n(t))$, which can be considered as a *free deconvolution* problem (see (2)).

2 Main results

For $g \in L^1(\mathbb{R})$, let $g^\star$ denote the Fourier transform of g.

2.1 Free deconvolution by subordination method

Let μ be a probability measure on $\mathbb{R}$, the Cauchy transform of μ is defined by $G_\mu(z) = \int_{\mathbb{R}} \frac{\mathrm{d}\mu(x)}{z-x}$, $z \in \mathbb{C}\backslash\mathbb{R}$. Denote the upper half-plane $\mathbb{C}^+ := \{z \in \mathbb{C} | \mathrm{Im}(z) > 0\}$ and $\mathbb{C}_\gamma := \{z \in \mathbb{C}^+ | \mathrm{Im}(z) > \gamma\}$, for $\gamma > 0$. We also can define $F_\mu(z) := 1/G_\mu(z)$, for $z \in \mathbb{C}^+$. Remind that $\mu_t = \mu_0 \boxplus \sigma_t$ (see (2)), we can prove that there exist unique subordination functions w_1 and w_{fp} from $\mathbb{C}_{2\sqrt{t}}$ onto $\mathbb{C}^+$, satisfying some specific properties, in particular:

$$F_{\mu_0}(z) = F_{\sigma_t}(w_1(z)) = F_{\mu_t}(w_{fp}(z)), \text{ for } z \in \mathbb{C}_{2\sqrt{t}}.$$

Moreover, set $K_z(w) := t.G_{\sigma_t}\Big(w + F_{\mu_t}(w) - z\Big) + z$ for $w \in \mathbb{C}_{\frac{1}{2}\mathrm{Im}(z)}$. Then, for any $z \in \mathbb{C}_{2\sqrt{t}}$ we have $K_z\big(w_{fp}(z)\big) = w_{fp}(z)$, and for any $w \in \mathbb{C}_{\frac{1}{2}\mathrm{Im}(z)}$, $K_z^{\circ m}(w)$ converges to $w_{fp}(z)$ when $m \to +\infty$. These results on the subordination function w_{fp} are adapted from [2], but in the special case of the free deconvolution by the semi-circular distribution σ_t.
As a consequence, we establish a fixed-point equation which allows us to obtain the Cauchy transform of the initial condition μ_0 from the fixed-point function w_{fp} as

$$G_{\mu_0}(z) = G_{\mu_t}\big(w_{fp}(z)\big) = \frac{1}{t}\big(w_{fp}(z) - z\big), z \in \mathbb{C}_{2\sqrt{t}}. \tag{3}$$

Let $\mathcal{C}_\gamma$ denote the centered Cauchy distribution with parameter $\gamma > 0$, and $*$ denote the classical convolution operator. Then, using the Stieltjes-inversion-formula, we get:

$$f_{\mu_0 * \mathcal{C}_\gamma}(x) = \frac{1}{\pi t}\big[\gamma - \mathrm{Im} w_{fp}(x + i.\gamma)\big], \text{ for } x \in \mathbb{R}, \gamma > 2\sqrt{t}. \tag{4}$$

2.2 Construction of statistical estimator of p_0

In order to construct an estimator of p_0, the first-step is to estimate $w_{fp}(z)$ for $z \in \mathbb{C}_{2\sqrt{t}}$ using equation (3). For this reason, it is natural to replace μ_t by its empirical measure μ_t^n. More precisely, remind that we do not observe directly the measure μ_t, but the matrix $X^n(t)$ for a given n. Then, for $z \in \mathbb{C}^+$, in equation (3) replacing $G_{\mu_t}(z)$ by:

$$\widehat{G}_{\mu_t^n}(z) := \int_{\mathbb{R}} \frac{d\mu_t^n(\lambda)}{z-\lambda} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{z-\lambda_j^n(t)} = \frac{1}{n}\mathrm{tr}\Big(\big(zI_n - X^n(t)\big)^{-1}\Big),$$

with I_n the identity matrix, leads to the following theorem:

Theorem-Definition 2.1 *There exists a unique fixed point to the following functional equation in $w(z)$: $\frac{1}{t}(w(z)-z) = \widehat{G}_{\mu_t^n}(w(z))$, for $z \in \mathbb{C}_{2\sqrt{t}}$. This fixed-point is denoted by $\widehat{w}_{fp}^n(z)$. Moreover, $\widehat{w}_{fp}^n(z) \in \mathbb{C}_{\frac{1}{2}Im(z)}$ and $\big|\widehat{w}_{fp}^n(z) - z\big| \leq \sqrt{t}$.*

Now, using (4), in order to finally obtain an estimator of p_0, the second-step is to make a classical deconvolution by $\mathcal{C}_\gamma$. Hence, we shift to Fourier-transform domain. Recall that for the Cauchy distribution $\mathcal{C}_\alpha$ with $\alpha > 0$, $f_\alpha^\star(\xi) = e^{-\alpha|\xi|}$ for $\xi \in \mathbb{R}$. We now define our ultimate estimator for the density function p_0 from its Fourier transform:

Definition 2.2 *For $\gamma > 2\sqrt{t}$, consider a bandwidth $h > 0$ and the sinc kernel K, namely $K(x) = sinc(x) = \sin(x)/(\pi x)$, with $K_h(x) := \frac{1}{h}.K\big(\frac{x}{h}\big)$ for $x \in \mathbb{R}$. We define the estimator $\widehat{p}_{0,h}$ of p_0 by its Fourier transform:*

$$\widehat{p}_{0,h}^\star(z) = e^{\gamma.|z|}.K_h^\star(z).\frac{1}{\pi t}\left[\gamma - \big(Im\widehat{w}_{fp}^n(\cdot + i.\gamma)\big)^\star(z)\right]. \tag{5}$$

From (5), we can see that the behavior of $\widehat{p}_{0,h}$ depends heavily on properties of $\widehat{w}_{fp}^n$. Thus, we have proved first the consistency of $\widehat{w}_{fp}^n$ in the following proposition, which is crucial to study the convergence of $\widehat{p}_{0,h}$.

Proposition 2.3 *Let $\gamma > 2\sqrt{t}$. Suppose p_0 satisfies $\int_{\mathbb{R}} \log(x^2+1)p_0(x)dx < +\infty$. Then:*
(i) For any $z \in \mathbb{C}_{2\sqrt{t}}$, the estimator $\widehat{w}_{fp}^n(z)$ converges almost surely to $w_{fp}(z)$ as $n \to \infty$.
(ii) The convergence is uniform on $\mathbb{C}_\gamma$.
(iii) The convergence rate on $\mathbb{C}_\gamma$: $\sup_{n\in\mathbb{N}} \sup_{z\in\mathbb{C}_\gamma} \mathbb{E}\left[n\big|\widehat{w}_{fp}^n(z) - w_{fp}(z)\big|^2\right] < +\infty$.

2.3 Mean integrated squared error (MISE)

By Parseval's equality, we consider the classical bias-variance decomposition of the L^2-risk:

$$\|\widehat{p}_{0,h} - p_0\|_{L^2(\mathbb{R})}^2 = \frac{1}{2\pi}\left\|\widehat{p}_{0,h}^\star - p_0^\star\right\|_{L^2(\mathbb{R})}^2 \leq \frac{1}{\pi}\left\|\widehat{p}_{0,h}^\star - K_h^\star.p_0^\star\right\|_{L^2(\mathbb{R})}^2 + \frac{1}{\pi}\left\|K_h^\star.p_0^\star - p_0^\star\right\|_{L^2(\mathbb{R})}^2.$$

Actually, the study for the variance term is quite involved and the order is provided by the following theorem.

Theorem 2.4 *Assume that there exists a constant $C > 0$ such that*

$$\mu_0\big((\kappa, +\infty)\big) \leq \frac{C}{\kappa}, \text{ for sufficiently large } \kappa > 0. \tag{6}$$

Then, for any $\gamma > 2\sqrt{t}$, there exists a constant $C_{var}(t)$ only depending on t such that for any $h > 0$ and n large enough,

$$\mathbb{E}\Big(\big\|\widehat{p}^{\star}_{0,h} - K^{\star}_h p^{\star}_0\big\|^2_{L^2(\mathbb{R})} \Big) \leq \frac{\gamma^8}{(\gamma^2 - 4t)^4}.\frac{C_{var}(t).e^{\frac{2\gamma}{h}}}{n}.$$

We consider first the space $\mathcal{S}_s(a, r, L)$ of supersmooth densities defined for $a > 0$, $L > 0$ and $r > 0$ by: $\mathcal{S}_s(a, r, L) = \big\{g \text{ density such that } \int_{\mathbb{R}} |g^{\star}(\xi)|^2.e^{2a|\xi|^r} \mathrm{d}\xi \leq L\big\}$.
For the bias, we classically have for $p_0 \in \mathcal{S}_s(a, r, L)$: $\|K^{\star}_h.p^{\star}_0 - p^{\star}_0\|^2_{L^2(\mathbb{R})} \leq Le^{-2ah^{-r}}$. We therefore obtain:

$$MISE := \mathbb{E}\Big[\big\|\widehat{p}_{0,h} - p_0\big\|^2_{L^2(\mathbb{R})} \Big] \leq Le^{-2ah^{-r}} + \frac{\gamma^8}{(\gamma^2 - 4t)^4}.\frac{C_{\mathrm{var}}(t).e^{\frac{2\gamma}{h}}}{n}. \tag{7}$$

The rates of convergence are summed up in the following corollary, adapted from the computation of [6]. One can see that there are three cases to consider to derive rates of convergence: $r = 1$, $r < 1$ and $r > 1$, depending on which the bias or variance term dominates the other.

Corollary 2.5 *Suppose that μ_0 satisfies Assumption* (6) *and the density p_0 belongs to the space $\mathcal{S}_s(a, r, L)$ for $a > 0$, $r > 0$ and $L > 0$. Then, for any $\gamma > 2\sqrt{t}$, we have:*

$$\mathbb{E}\Big[\|\widehat{p}_{0,h} - p_0\|^2_{L^2(\mathbb{R})}\Big] = \begin{cases} O\big(n^{-\frac{a}{a+\gamma}}\big) & \text{if } r = 1 \\ O\Big(\exp\Big\{ -\frac{2a}{(2\gamma)^r}\Big[\log n + (r-1)\log\log n + \sum_{i=0}^{k} b^*_i (\log n)^{r+i(r-1)} \Big]^r \Big\}\Big) & \text{if } r < 1 \\ O\Big(\frac{1}{n} \exp\Big\{ \frac{2\gamma}{(2a)^{1/r}}\Big[\log n + \frac{r-1}{r}\log\log n + \sum_{i=0}^{k} d^*_i (\log n)^{\frac{1}{r} - i\frac{r-1}{r}} \Big]^{1/r} \Big\}\Big) & \text{if } r > 1, \end{cases}$$

where $k \in \mathbb{N}$ is such that $\frac{k}{k+1} < \min\big(r, \frac{1}{r}\big) \leq \frac{k+1}{k+2}$, and where the constants b^*_i and d^*_i are computable.

Now, let us consider Sobolev ordinary-smooth type regularities. Assume that p_0 belongs to the Sobolev class $\mathcal{S}_b(\beta, L)$ defined for $\beta > 0$ and $L > 0$ as:

$$\mathcal{S}_b(\beta, L) = \Big\{ g \text{ density such that } \int_{\mathbb{R}} |g^{\star}(\xi)|^2.(1 + \xi^2)^{\beta} d\xi \leq L \Big\}.$$

We have then for the bias term: $\|K^{\star}_h.p^{\star}_0 - p^{\star}_0\|^2_{L^2(\mathbb{R})} \leq L.h^{2\beta}$. Furthermore, using Theorem 2.4, we obtain the following result.

Corollary 2.6 *Suppose that μ_0 satisfies Assumption* (6) *and the density p_0 belongs to the space $\mathcal{S}_b(\beta, L)$ for $\beta > 0$ and $L > 0$. Then, for any $\gamma > 2\sqrt{t}$ and by choosing the bandwidth $h = \widetilde{C}.\log^{-1}(n)$ with $\widetilde{C} > 2\gamma$, we have:*

$$\mathbb{E}\Big[\, \|\widehat{p}_{0,h} - p_0\|^2_{L^2(\mathbb{R})} \,\Big] = O\Big((\log n)^{-2\beta}\Big).$$

Now, let us discuss the optimality of the convergence rates stated in Corollaries 2.5 and 2.6. It is relevant to connect them with the minimax rates obtained in the classical statistical density deconvolution problem by Butucea and Tsybakov in [3] for supersmooth densities or in Fan and Koo [5] for Sobolev ordinary-smooth type regularities. Here, our estimation strategy converts the initial free deconvolution problem into the classical deconvolution problem (4) between μ_0 and the Cauchy distribution $\mathcal{C}_\gamma$. Thus, our observation scheme is more intricate and involved than the framework of classical density deconvolution tackled in [3] and [5]. If our observations had been distributed according to the density $f_{\mu_0 \star \mathcal{C}_\gamma}$ as in [3] and [5], for a given γ, the upper bound of the variance term given by Theorem 2.4 as well as the bounds for the bias mentioned above would have been optimal. Consequently, as part of our strategy, we expect that our rates of convergence cannot be improved for a given γ.

References

[1] G.W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[2] O. Arizmendi, P. Tarrago, and C. Vargas. Subordination methods for free deconvolution. *Ann. Inst. H. Poincaré Probab. Statist.*, 56(4):2565–2594, 2020.

[3] C. Butucea and A. B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. I. *Teor. Veroyatn. Primen.*, 52(1):111–128, 2007.

[4] S. Dallaporta and M. Février. Fluctuations of linear spectral statistics of deformed Wigner matrices. submitted. `hal-02079313`, 2019.

[5] J. Fan and J.-Y. Koo. Wavelet deconvolution. *IEEE Trans. Inform. Theory*, 48(3):734-747, 2002.

[6] C. Lacour. Rates of convergence for nonparametric deconvolution. *Comptes rendus de l'Académie des sciences. Série I, Mathématique*, 342(11):877-882, 2006.

[7] M. Maïda, TD. Nguyen, TM. Pham Ngoc, V. Rivoirard and VC. Tran. Statistical deconvolution of the free Fokker-Planck equation at fixed time. submitted. `hal-02876999`, 2020.

Full Draft as of 26/02/2021

Conditional Kaplan-Meier survival function: Illustration for female and male promotion in Science

Jacques Mairesse[a] and Michele Pezzoni[b]

Abstract: In event history analysis, the standard practice is to compute the 'Kaplan-Meier' survival function and represent it graphically as a first exploratory analysis, followed by a Cox's regression with a number of control variables and assuming either a proportional or fully parametrized hazard function. In the present contribution, we show that it is also possible to compute and represent graphically a 'Kaplan-Meier' survival function conditionnal on all the control variables. One would think that it is particularly interesting to assess the results of an event history analysis and present them graphically by comparing the conditional Kaplan-Meier (CKM) survival function to the usual Kaplan-Meier (UKM) survival function. However this as never been done to our knowledge. One reason is probably that it is not trivial to implement ~~a conditional Kaplan-Meier (~~CKM) survival function. One needs to rely on a fully parametrized hazard function, while a majority of studies rely on the less specific and more convenient proportional hazard function.

To illustrate the interest of the (CKM) survival function as a tool, we consider what would have been its application in a study where we have investigated what are the factors of the promotion of female and male scientists at the French Institute of Physics (INP) at CNRS

Keywords: Event history analysis; Hazard rate function; Kaplan-Meier' survival function; Fully parametric event history analysis.

[a]CREST ENSAE (France); UNU-MERIT, Maastricht University (Netherlands); EHESS (France); NBER (USA); email: *mairesse@ensae.fr*.

[b]GREDEG, CNRS, Université Côte d'Azur (France); OST, HCERES, (France); ICRIOS, Bocconi University (Italy); email: *michele.pezzoni@unice.fr*.

Acknowledgments: We are grateful to Fabiana Visentin for her collaboration in two companion papers. We are also indebted to Alain Schuhl and Anne Sigogneau for their great help in implementing an online survey on family and research responsibilities of the female and male physicists of the Institute of Physics (INP) at CNRS.

1. Introduction

In event history analysis, the standard practice is to compute the 'Kaplan-Meier' survival function and represent it graphically as a first exploratory analysis, followed by a Cox's regression with a number of control variables and assuming either a proportional or fully parametrized hazard function. In the present contribution, we show that it it is also possible to compute and represent graphically a 'Kaplan-Meier' survival function conditionnal on all the control variables.. To do so, one needs to rely on a fully parametrized hazard function, while a majority of studies rely on the less specific and more convenient proportional hazard function.

To illustrate the interest of the conditionl survival function as a tool, we consider what would have been its application in a study where we have investigated what are the factors of the promotion of female and male scientists at the French Institute of Physics (INP) at CNRS

Section 2 of our paper explain in details how to defin and implement conditional Kaplan-Meier survival function. Section 3 present an illustration of conditionl survival function as a tool in the case of event analysis of female and male physicists promotion at CNRS, that we have published in two compagnion studies to the present paper: Mairesse-Pezzoni-Visentin, 2019 and 2020.

2. Defining and implementing the conditional Kaplan-Meier survival function

The Kaplan-Mayer statistic is often used to estimate the survival function in presence of individual survival data. The statistic for the first survival period ($\hat{S}(t_1)$) is calculated as 1 minus the ratio between the number of individuals who experience the event in the first period (d_1) over the number of individuals at risk (n_1). The statistic for the second survival period ($\hat{S}(t_2)$) is calculated as $\hat{S}(t_1)$ multiplied by 1 minus the ratio between the number of individuals who experience the event in the second period (d_2) over the number of individuals at risk (n_2). For a generic survival period *t*, the Kaplan-Mayer statistic is calculated as in Equation 1.

$$\hat{S}(t) = \prod_{t_i<t}(1 - \frac{d_i}{n_i}) \quad \text{Equation 1}$$

The Kaplan-Mayer statistic is often used as a first descriptive in many empirical exercises. The graphical representation of the Kaplan-Mayer statistic is a curve showing a series of decreasing "steps" resulting from considering discrete time periods to estimate the survival function, rather than considering continuous time.

Although its utility for descriptive purposes, the Kaplan-Mayer statistic is not calculated conditional on the individuals' characteristics. To calculate the survival functions controlling for individuals characteristics, we need to use a model from which we can estimate a conditional survival function. To do so, we proceed in three steps. First, we define a parametric model in which we predict the hazard rate conditional on the individuals' characteristics. Second, we show the relationship between the hazard rate and the survival function. Third, using the estimated parameters of the model predicting the hazard rate, we represent graphically the conditional survival function.

Parametric models used to predict the hazard rate

Parametric models predicting the hazard rate conditional on the individuals' characteristics can be generally represented as in Equation 2, where $h(t)$ is the hazard rate at time *t, x* is a vector of individual characteristicsβ_0, is the model constant, β is the vector of coefficients to be estimated, and $h_0(t)$ is the baseline hazard function.

$$h(t) = h_0(t)e^{(\beta_0 + x\beta)} \qquad \text{Equation 2}$$

The baseline hazard function represents the direct relationships between the time and the hazard rate. In other words, it represents the idea that the hazard rate is expected to increase with time despite the individuals' characteristics. In parametric models, the main problem is that the functional form of $h_0(t)$ is unknown. Often statisticians use a Weibull functional form to define the baseline hazard rate. Assuming that our baseline hazard function has a Weibull distribution $h_0(t) = \gamma\alpha t^{\alpha-1}$, we can rewrite Equation 2 as the following Equation 3,:

$$h(t) = \gamma\alpha t^{\alpha-1}e^{(x\beta)} \qquad \text{Equation 3.}$$

where $\gamma = e^{(\beta_0)}$, and the parameters α, β_0, β and γ are estimated using maximum likelihood.

Relationship between the hazard rate and the survival function

The hazard rate is defined as the ratio between the probability that the event occurs at time *t* (f(*t*)) and the probability that the event has not yet occurred until *t*, i.e., one minus the cumulative distribution function (F(t)), or Equation 4:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

The Weibull functional form has mathematical properties that allow to conveniently define the hazard rate ($h(t)$) as in Equation 5, the probability density function (f(*t*)) as in Equation 6, the cumulative distribution function (F(*t*)) as in Equation 7, and the survival function (1-F(*t*)) as in Equation 8. Specifically, the survival function can be written as in Equation 8, and the parameters γ and α derived from the estimations of Equation 3;

$$h(t) = \gamma\alpha t^{\alpha-1} \qquad \text{Equation 5}$$

$$f(t) = \gamma\alpha t^{\alpha-1} e^{-\gamma t^{\alpha}} \qquad \text{Equation 6}$$

$$F(t) = 1 - e^{-\gamma t^{\alpha}} \qquad \text{Equation 7}$$

$$S(t) = 1 - F(t) = -e^{-\gamma t^{\alpha}} \qquad \text{Equation 8}$$

Using the estimated parameters of the model we calculate the conditional survival function

Combining the results of the parameter estimates in Equation 3 and the mathematical relationships between the hazard rate (Equation 5) and the survival function (Equation 8) and assuming a Weibull distribution, we can predict the survival function at any period *t* conditional on the control variables (the vector *x* in Equation 3). This allows us to compare the graphical representation of the Kaplan- Meier survival function with the graphical representation of the predicted survival function conditional on control variables. Moreover, having assumed a Weibull distribution, allows us to consider continuous time, and represent graphically the survival function conditional on control variables as a continuous curve, while the usual Kaplan-Meier representation of the survival function is defined in discrete-time and itshe graphical representation is characterized by “steps”.

3. Illustration for an event analysis of female and male physicists promotion

at CNRS

3.1 Organizational setting, variables and descriptive statistics

The illustration we propose is based on a panel data sample of female and male researchers working at the Institute of Physics (INP), the Institute of CNRS specialized in the field of physics, for which we were able to combine administrative and bibliometric data, as well as specific information gathered from an online survey. Based on these data and relying on event history analysis, as explained in Section 2, we analyze promotions of these researchers from the entry positions or ranks of '*Chargé de Recherche'* to the highest rank of '*Directeur de Recherche'*, respectively denoted (CR) and (DR).

In this setting, in the two compagnion studies to the present analysis, we have investigated whether factors of promotion, direct and indirect, such as research output and family characteristics, can account for differences in the promotion rates of the female and male INP physicists from CR to DR ranks. In addition to the usual measure of numbers of publications and citations received, we have also considered complementary research activities such as mentoring, professional networking, fundraising, technology transfer, and project management activities. We have also considered family characteristics which we had retrieved in an online survey of INP researchers about the number and age of their children. We focus here on the impact of these family characteristics on the female and male physicists CR to DR promotion rates, as precisely proxied by the scientist's number of children until year (*t-1)*, and a dummy variable that equals one if the scientist has one child born in the last three years, zero otherwise, respectively denoted by Family size in year (t-1) and One child less than 3 years old.

Our study sample is an unbalanced panel of 7,805 observations for 604 INP scientists, of which 139 are women (23.0%), and 465 are men (77.0%). Each of them is an active researcher from when entering at CNRS until 2017, the year of our online survey and last year of our study sample. Table 1 shows that overall, during our study period, 276 (45.7%) researchers are promoted DR after 14.3 years, on average, and 328 (54.3%) researchers stay CR without being promoted to DR for 11.7 years, on average.

We see also in Table 1 that 56 (40.3%) female physicists are promoted DR after 15.8 years and their 220 (47.3%) male colleagues after 14.0 years. The difference of years to be promoted

from CR to DR for female and male physicists is of 1.8 years (=15.8-14.0), a rather small number, but statistically highly significant (the P-value of the test between the two means equals 0.0076). Relatedly, 83 (59.7 %) female researchers remain CR for 13.3 years until the end of the study period, while 245 (62.7 %) male researchers remain CR for 11.2 years

Table 1: Promotion of INP scientists from CR to DR

Numbers	Scientists			Scientist-year Observations			Average number of years at risk of promotion		
	All	Female	Male	All	Female	Male	All	Female	Male
Recruited CR	604	139	465	7805	1989	5816	12.9	14.3	12.5
Promoted DR	276	56	220	3967	888	3 079	14.3	15.8	14.0
Not promoted DR	328	83	245	3838	1101	2737	11.7	13.3	11.2
Share of promoted (%)	45.7	40.3	47.3	50.8	44.6	52.9	--	--	--

Note: For the individuals not promoted to DR, the average year duration is right-censored.

.

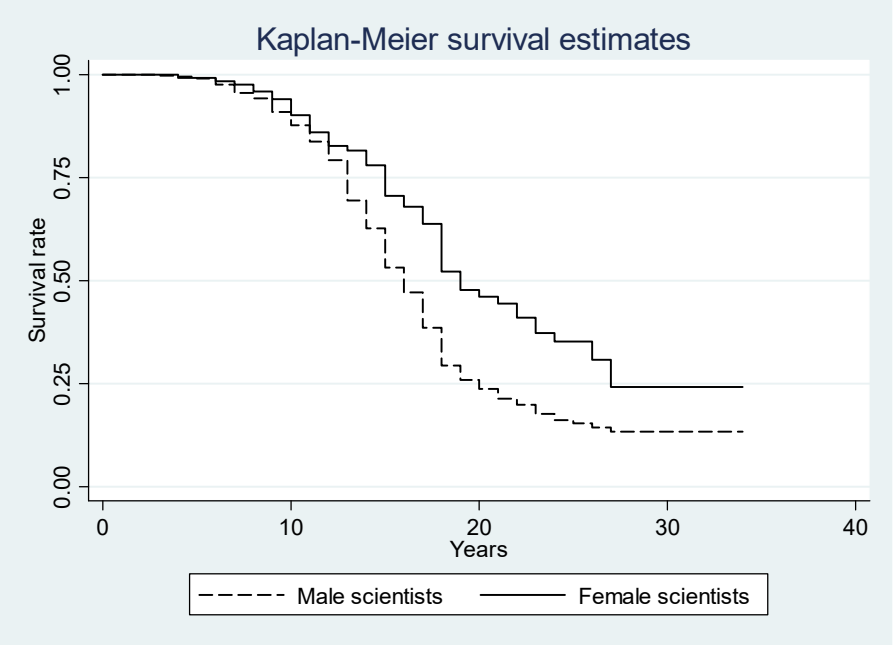
In the Appendix, we document in Table A1 the descriptive statistics for the time-invariant and time-variant variables. The descriptive statistics for the time-invariant variables are calculated over the sample of 604 researchers. Statistics for the time-variant variables are calculated over the 7,805 periods when the 604 researchers are at risk of promotion, namely from the year of entry at CNRS until the year of promotion to DR, or until 2017 if a researcher remains CR. We observe that the differences between female and male researchers are substantial and significant, as they are between researchers promoted and not-promoted to DR. We also see that the averages of the time-invariant covariates often small and not statistically significant. As concerns family characteristics the family size and the child less than three-years old dummy are weakly statistically different between female and male scientists, and on average respectively equal 1.19% and 26% for female and 1.13% and 28% for male.

3.2 Results

In Figure 1 we show the Kaplan-Meier non-parametric survival function for female and male scientists separately, when we follow the standard practice of computing and representing the 'Kaplan-Meier' survival function graphically, as a first exploratory analysis. We see clearly that the average survival rate for female physicists as CR(i.e., the time elapsed from the recruitment

as CR to the promotion to DR) is higher than for males in all periods, meaning that the average promotion rate to DR is lower for female than males. If we compute the log-rank test of equality between the two Kaplan-Meir survival functions, we also find that it is rejected with a P-value of 0.0004 (Chi2 =12.51).

Figure 1: Kaplan-Meier survival function estimates by gender



Note: The x-axis years correspond to the time elapsed from the recruitment as CR to the promotion to DR.

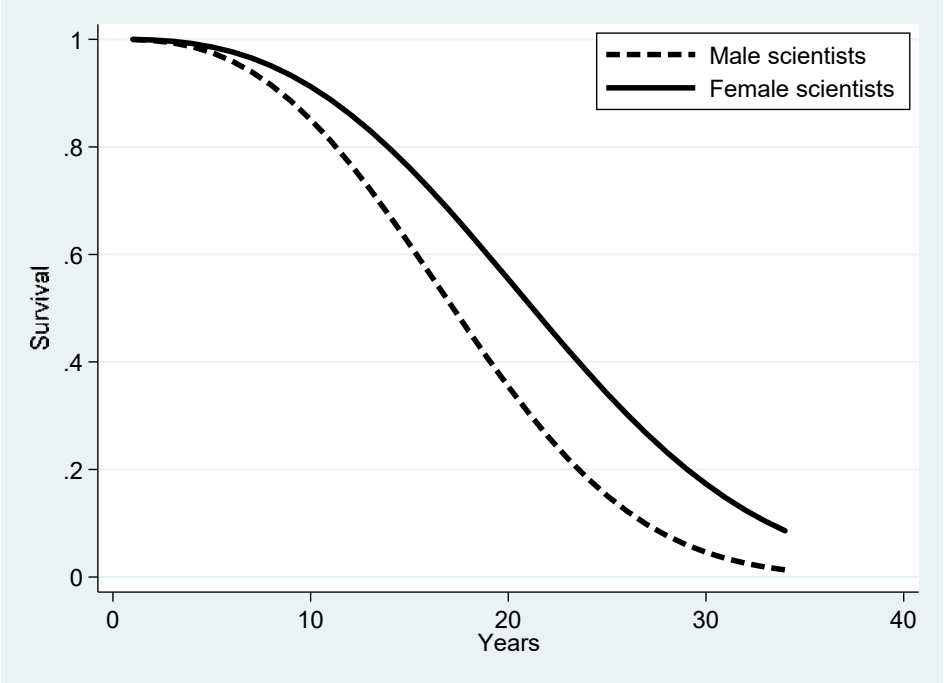
Table A2 in the Appendix reports the maximum likelihood estimates of the parametric proportional hazard model, assuming a Weibull distribution as in Equation 3. Figure 2 shows the predicted survival function according to the estimates in Column 1, Figure 3 the predicted survival function according to the estimates in Column 2, and Figure 4 the predicted survival function according to the estimates in Column 3.

The graphical representation in Figure 2 of the survival functions based on our parametric model estimates including only the *Female* dummy variable is very similar the one of Figure 1 and the estimates derived from the parametric model including only the *Female* dummy variable (Figure 2). In both cases, the curve representing the survival function of female scientists is above the curve of male scientists. This means that, for a given value of survival time, female scientists

are more likely than male scientists to be in the CR career step (or, equivalently, are less likely to be promoted DR).

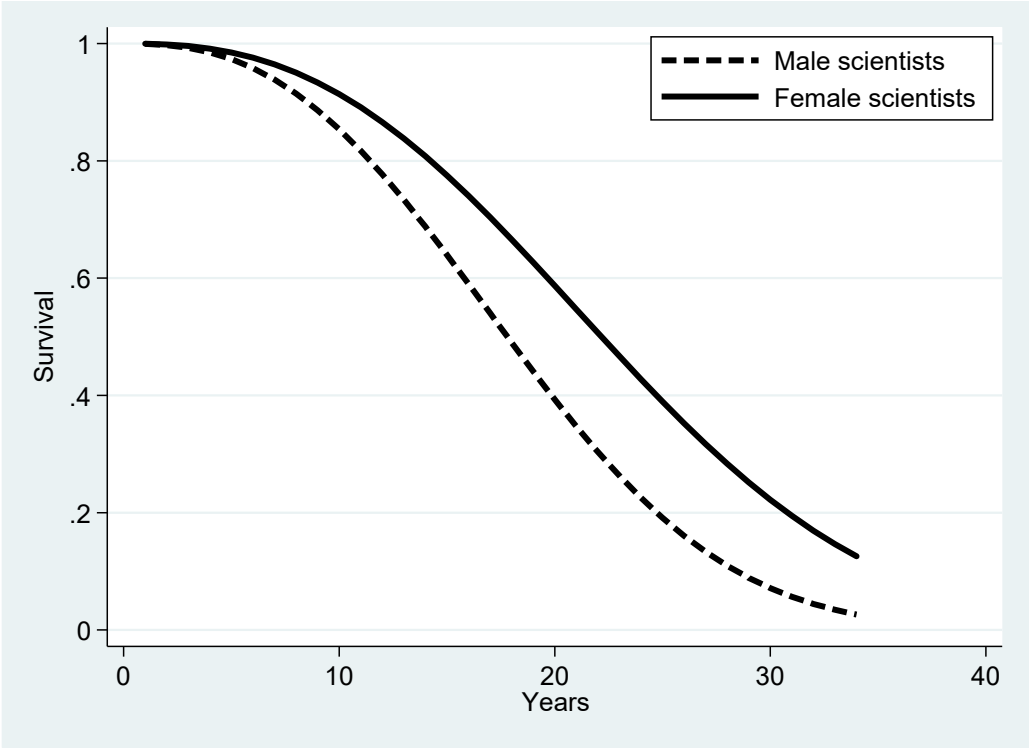
When we control only for family characteristics, we have a similar conclusion.The female scientists' estimated survival function is still above the male survival function meaning that controlling for family characteristics does not affect the female likelihood of promotion to DR. However, when we control for the academic characteristics of scientists (i.e., scientific outputs, fundraising ability, and teaching experience, etc.), we find interestingly two survival functions that are not significantly different for female and male scientists.

Figure 2: Weibull survival distribution by gender fitted from estimates in Table 1, column 1.



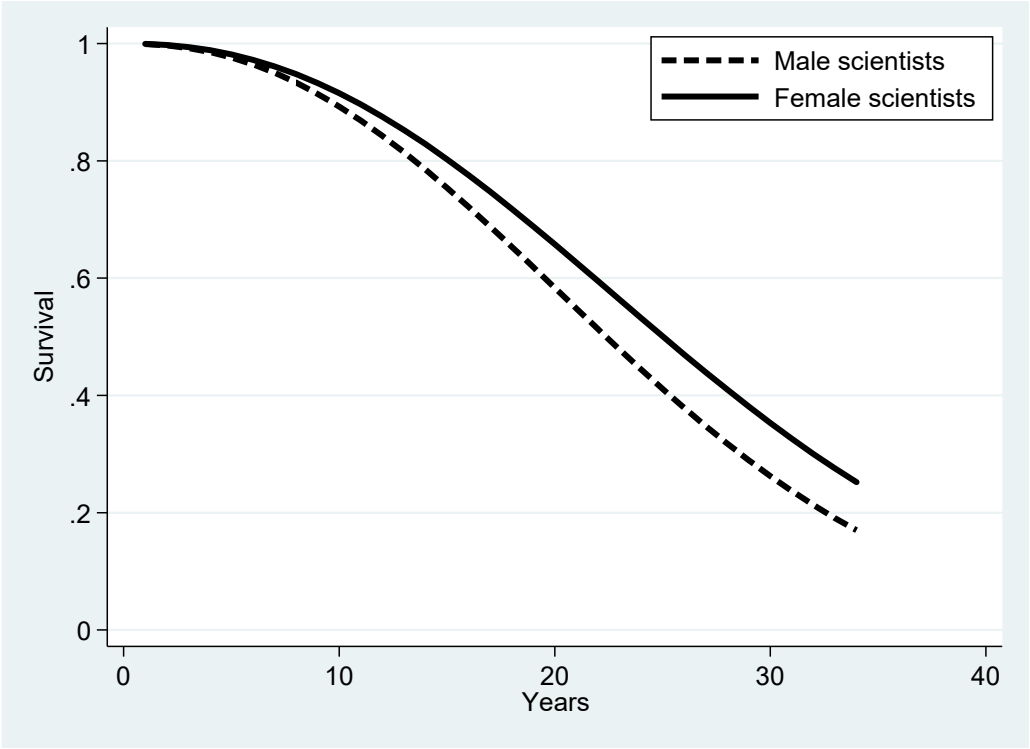
Note: The analysis time corresponds to the number of years elapsed from the recruitment as CR to the promotion to DR.

Figure 3: Weibull survival distribution by gender fitted from estimates in Table 1, column 2.



Note: The analysis time corresponds to the number of years elapsed from the recruitment as CR to the promotion to DR.

Figure 4: Weibull survival distribution by gender fitted from estimates in Table 1, column 3.



Note: The analysis time corresponds to the number of years elapsed from the recruitment as CR to the promotion to DR.

4. Conclusion

This paper aims to compare different ways of estimating the survival function for a longitudinal sample of individuals at risk of an event. We estimate the survival function following three different approaches. The first approach is the non-parametric Kaplan-Meier estimate, the second approach calculates the survival function relying on the estimates of a parametric model including only one variable of interest and no controls, and the third approach calculates the survival function relying on the estimates of a parametric model including the variable of interest and a set of controls. For each approach, we produce a graphical representation of the survival function.

The three graphical representation of the survival functions shows some peculiarities. The survival curve derived from the non-parametric Kaplan-Meier approach is calculated without assuming any functional form but it requires dividing the survival time into discrete time intervals. Moreover, it does not allow to calculate confidence intervals and to control for individual characteristics. Different from the Kaplan-Meier approach, the parametric model including only a variable of interest allows to consider continuous survival time and to include confidence intervals in the graphical representation. Finally, the survival function relying on the estimates of a parametric model including a set of controls allows calculating the survival function and its confidence intervals conditioning on the individual characteristics.

We suggest that a good practice in the event history analyses would be to compare systematically the three graphics resulting from the three ways of estimating the survival function. We illustrate our approach in an empirical exercise comparing three graphics of the survival functions of female and male researchers at risk of promotion.

References

Allison, Paul David. [1984]. *Event History Analysis Regression for Longitudinal Event Data*. Beverly Hills, Calif.: Sage Publications.

Cox, David R. [1972]. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.

Kaminski, D. and C. Geisler. [2012]. "Survival Analysis of Faculty Retention in Science and Engineering by Gender." *Science* 335(6070):864–66.

Long, J. Scott, Paul D. Allison, and Robert McGinnis. [1993]. "Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity." *American Sociological Review* 58(5):703–722.

Mairesse Jacques and Michele Pezzoni. [2015]. "Does Gender Affect Scientific Productivity? A Critical Review of the Empirical Evidence and a Panel Data Econometric Analysis for French Physicists". *Revue Economique,* 66(1), p.65-113.

Mairesse, Jacques, Michele Pezzoni, and Fabiana Visentin. [2019]. "Impact of Family Characteristics on the Gender Publication Gap: Evidence for Physicists in France." *Interdisciplinary Science Reviews*, 44 (2): 204–220.

Mairesse, Jacques, Michele Pezzoni, and Fabiana Visentin. [2020]. "Does Gender Matter for Promotion in Science? Evidence from Physicists in France." *Revue économique* 71(6): 1093–1131.

Sabatier, Mareva. [2010]. "Do Female Researchers Face a Glass Ceiling in France? A Hazard Model of Promotions." *Applied Economics* 42(16):2053–62.

Wooldridge, Jeffrey M. [2002]. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Appendix A

Table A1: Descriptive statistics.

	(1) All	(2) Female	(3) Male	(4) P-val. 2-3	(5) Promoted to DR	(6) Not promoted	(7) P-val. 5-6
Time-invariant covariates							
604 researchers (139 females, 465 males)							
Female scientist	0.23	1	0		0.20	0.25	0.145
Male scientist*	0.77	0	1		0.80	0.75	0.145
Age at CR	31.25	31.1	31.3	0.464	30.41	31.97	0.000
Recruited as CR1	0.22	0.23	0.22	0.746	0.25	0.20	0.069
Ph.D. in a Paris university	0.35	0.29	0.37	0.105	0.35	0.35	0.955
Ph.D. in a French university	0.44	0.41	0.45	0.379	0.41	0.46	0.238
Ph.D. in a foreign university (ref.)	0.21	0.30	0.18	0.003	0.24	0.19	0.134
Ph.D. graduation year*	1996	1995	1997	0.253	1990	2002	0.000
Ph.D. graduation year 2001-2017	0.40	0.37	0.42	0.311	0.08	0.68	0.000
Ph.D. graduation year 1991-2000	0.29	0.27	0.29	0.629	0.36	0.23	0.001
Ph.D. graduation year before 1991 (ref.)	0.31	0.36	0.29	0.120	0.56	0.09	0.000
Section 2: Physical theories	0.19	0.10	0.21	0.003	0.21	0.17	0.312
Section 3: Condensed matter physics (structures and electronic properties)	0.24	0.21	0.25	0.349	0.22	0.25	0.467
Section 4: Atoms and molecules, optics and lasers, hot plasma physics	0.29	0.31	0.29	0.596	0.29	0.29	0.939
Section 5: Condensed matter physics (organizations and dynamics)	0.28	0.38	0.26	0.004	0.28	0.29	0.914
Time-variant covariates							
7805 periods at risk of promotion (1989 for females, 5816 for males)							
Cumulated number of articles in t-1	21.44	15.63	23.43	0.000	41.86	20.70	0.000
Cumulated number of conference papers in t-1	4.06	2.86	4.47	0.000	7.43	3.94	0.000
Average number of citations in t-1	1.82	1.23	2.016	0.000	1.88	1.81	0.612
Cumulated number of collaborators in t-1	25.02	19.58	26.88	0.000	43.50	24.34	0.000
Cumulated number Ph.D. theses supervised in t-1	0.16	0.13	0.17	0.002	0.69	0.14	0.000
Cumulated years as head of a research team in t-1	0.75	0.75	0.74	0.848	2.55	0.68	0.000
Cumulated years with other research responsibilities in t-1	0.42	0.28	0.47	0.000	1.08	0.39	0.000
Family size in t-1	1.14	1.19	1.13	0.042	1.79	1.12	0.000
One child less than 3 years old	0.28	0.26	0.28	0.038	0.17	0.28	0.000
At least one EPO patent in t-1	0.05	0.04	0.05	0.537	0.08	0.05	0.008
At least one ANR or EU grant in t-1	0.04	0.03	0.04	0.066	0.05	0.04	0.183
Gender parity initiative (MPPF)	0.38	0.33	0.40	0.000	0.19	0.39	0.000
Share of individuals with at least one child in the observation period*	0.75	0.79	0.73	0.168	0.81	0.70	0.001
Average number of publications at promotion time*	43.46	32.98	46.59	0.000	49.29	38.55	0.000

Note: Column 4 shows the P-value of the tests for mean equality between females (Column 2) and males (Column 3), while Column 7 shows the P-value of the tests for mean equality between researchers promoted to DR (Column 5) and not promoted (Column 6). The average values for the time-invariant covariates are calculated at the researcher level, while the time-variant covariates are calculated as an average for all the period at risk of promotion. *Average statistics mentioned in the text, but not entering in the econometric analysis.

Table A2: Event history analysis for promotion to DR using maximum likelihood estimations of the parametric model assuming the Weibull functional form of the baseline hazard ($h_0(t) = \gamma\alpha t^{\alpha-1}$).

	(1) All Hazard ratio	(2) All Hazard ratio	(3) All Hazard ratio
Female	0.569***	0.570***	0.778
	(0.100)	(0.0989)	(0.151)
Cumulated number of articles in t-1			1.027***
			(0.00487)
Cumulated number of conference papers in t-1			0.998
			(0.00741)
Average number of citations in t-1			1.069***
			(0.0245)
At least one EPO patent in t-1			1.208
			(0.299)
Cumulated number of collaborators in t-1			0.992**
			(0.00362)
Cumulated number Ph.D. theses supervised in t-1			1.233***
			(0.0855)
Cumulated years as head of a research team in t-1			1.088***
			(0.0226)
Cumulated years with other research responsibilities in t-1			0.973
			(0.0305)
At least one ANR or EU grant in t-1			0.808
			(0.279)
Family size in t-1		1.162***	1.132*
		(0.0664)	(0.0720)
One child less than 3 years old		0.806	0.820
		(0.143)	(0.146)
Age at CR			1.003
			(0.0385)
Ph.D. from a Paris university			0.680*
			(0.152)
Ph.D. from a French university			0.801
			(0.174)
Ph.D. graduation year 2001-2017			0.775
			(0.275)
Ph.D. graduation year 1991-2000			1.100
			(0.194)
Recruited as CR1			1.621**
			(0.387)
Gender parity initiative (MPPF)			0.835
			(0.208)
Constant	0.000335***	0.000387***	0.000492***
	(0.000110)	(0.000141)	(0.000609)
Section dummies	No	No	Yes
Log-likelihood	-313.747	-309.317	-226.885
Weibull parameter	2.683	2.563	2.244
Observations	7,805	7,805	7,805

Clustering Data with Nonignorable Missingness using Semi-Parametric Mixture Models

Matthieu Marbac[1] & Marie Du Roy de Chaumaray [2]

[1] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*
matthieu.marbac-lourdelle@ensai.fr
[2] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*
marie.du-roy-de-chaumaray@ensai.fr

Résumé. On s'intéresse au clustering de données continues sous un processus de génération de valeurs manquantes non ignorable. Le clustering est effectué par un mélange semi-paramétrique. L'estimation est faite par maximisation de la vraisemblance lissée au moyen d'un algorithme de Majoration-Minimisation. L'apport de notre approche est illustrée par sur données simulées.

Mots-clés. Clustering, Modèle de Mélange, Valeurs manquantes non ignorables, Vraisemblance lissée.

Abstract. We are concerned in clustering continuous data sets subject to nonignorable missingness. Clustering is achieved by a specific semi-parametric mixture. Estimation is performed by maximizing the smoothed likelihood via a Majoration-Minimization algorithm. Simulated data illustrates the benefits of the poroposed method.

Keywords. Clustering, Mixture Model, Nonignorable Missigness, Smoothed Likelihood.

1 Introduction

Mixture models permit to achieve the clustering purpose in a rigorous context but the case where data have missingness is generally neglected. Moreover, the missing not at random scenario (MNAR; Little and Rubin (2019)), where the missingness process depends on the missing values even conditional on the observed covariates, generally requires the missingness process to be considered to obtain consistent estimators. However, few statistical methods permit this scenario because the models are often not identifiable based on the observed data.

Two clustering approaches allow data subject to the MNAR scenario to be analyzed. Chi et al. (2016) introduces the K-POD algorithm that extends the K-means to the case of missing data even if the missing mechanism if unknown. However, this approach suffers from the standard drawbacks of the K-means algorithm (*i.e.,* assumptions of spherical clusters and equals proportions of the clusters). Alternatively, using a *selection model*

approach Miao et al. (2016) proposed a specific Gaussian mixtures and t-mixtures to analyze data under MNAR scenario. For such approach, the missingness process must be specified (probit and logit distributions are generally used). However, this approach produces strong bias if the parametric assumptions (made on the covariate distribution or on the missingness process) are violated.

In this paper, clustering is performed via a mixture model that uses a *pattern-mixture model* approach with non-parametric distributions. Thus, no assumptions are made on the data distribution or on the missingness process except that the variables are independent within components. Note that this assumption is quite standard for semi-parametric mixtures (Levine et al., 2011; Kasahara and Shimotsu, 2014). For each mixture component, we estimate, for each variable, its probability to be observed and its conditional distribution given the variables is observed. We emphasize that our concern is clustering and not imputation or density estimation. Indeed, without adding assumptions, the distribution of the variables within component cannot be estimated by our procedure. Estimation of mixture is done by maximizing the smoothed likelihood (Levine et al., 2011).

The paper is organized as follows. Section 2 introduces the model Section 3 focuses on the estimation. Section 4 illustrates the relevance of the approach. Section 5 gives a conclusion. More details are available in Du Roy de Chaumaray and Marbac (2020).

2 Mixture for nonignorable missingness

2.1 The data

The observed sample is composed of n independent and identically distributed subjects arisen form K homogeneous subpopulations. Each subject is described by d continuous variables and some realizations of these variables may be unobserved. The probability, for a variable, to be not observed is allowed to depend on the values of the variable itself and the subpopulation membership.

Each subject i is described by a vector of three variables $(\boldsymbol{X}_i^\top, \boldsymbol{R}_i^\top, \boldsymbol{Z}_i^\top)^\top$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top \in \mathbb{R}^d$ is set of continuous variables, $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{id})^\top \in \{0,1\}^d$ indicates whether X_{ij} is observed ($R_{ij} = 1$) and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})^\top$ indicates the subpopulation of subject i ($Z_{ik} = 1$ if subject i belongs to subpopulation k and otherwise $Z_{ik} = 0$). Each subject belongs to one subpopulation such that $\sum_{k=1}^K Z_{ik} = 1$. The realizations of $\boldsymbol{Z}_i$ are unobserved and a part of the realizations of $\boldsymbol{X}_i$ can be unobserved too. Therefore, the observed variables for subject i are $(\boldsymbol{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$ where $\boldsymbol{X}_i^{\text{obs}}$ is composed of the elements of $\boldsymbol{X}_i$ such that $R_{ij} = 1$ and the unobserved variables for subject i are $(\boldsymbol{X}_i^{\text{miss}\top}, \boldsymbol{Z}_i^\top)^\top$ where $\boldsymbol{X}_i^{\text{miss}}$ is composed of the elements of $\boldsymbol{X}_i$ such that $R_{ij} = 0$.

2.2 General mixture model

We use mixture models in a purpose of clustering and not for density estimation. Clustering aims to estimate the subpopulation memberships given the observed variables (*i.e.,* the realization of $\boldsymbol{Z}_i$ given $(\boldsymbol{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$) without assumption on the missingness process (*i.e.,* no assumption are made on the conditional distribution of $\boldsymbol{R}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i$). The probability distribution function (pdf) of $(\boldsymbol{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ for subpopulation k (*i.e.,* $Z_{ik} = 1$) is denoted by $g_k(\cdot)$. Thus, the pdf $(\boldsymbol{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ is defined by the pdf of a K-component mixture

$$g(\boldsymbol{x}_i, \boldsymbol{r}_i) = \sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}_i, \boldsymbol{r}_i), \tag{1}$$

where $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$ and $g_k(\cdot; \boldsymbol{\theta})$ is pdf of component k. From (1), using the *pattern-mixture model*, the pdf of component k is given by

$$g_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = g_k(\boldsymbol{r}_i) g_k(\boldsymbol{x}_i \mid \boldsymbol{r}_i). \tag{2}$$

For clustering, the *pattern-mixture model* should be preferred to *selection model* because it does not require to specify the missingness process, allows this process to be nonignorable and permits to easily obtain the conditional probabilities of the subpopulation membership given the distribution of the observed values

$$\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) = \frac{g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)}{\sum_{\ell=1}^K \pi_\ell g_\ell(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i \boldsymbol{\theta})}.$$

Indeed, integrating the pdf of component k over the missing variables $\boldsymbol{X}_i^{\text{miss}}$, we have

$$g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) = g_k(\boldsymbol{r}_i) g_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i).$$

Note that this approach does not permit to estimate the marginal distribution of $\boldsymbol{X}_i \mid \boldsymbol{Z}_i$ without adding assumptions on the missing process. Thus, the proposed approach can be used for clustering but not for density estimation.

2.3 Semi-parametric mixture for nonignorable missingness

A wide range of literature focuses on models assuming that conditionally on knowing the particular subpopulation the subject i came from, its coordinates $\boldsymbol{X}_i$ are independent. Thus, we extend this model for nonignorable missingness. The couples of variables $(X_{ij}, R_{ij})^\top$ are assumed to be conditionally independent given $\boldsymbol{Z}_i$. Thus, the distribution of $\boldsymbol{R}_i \mid \boldsymbol{Z}_i$ is a product of Bernoulli distributions and the conditional density of $\boldsymbol{X}_i \mid \boldsymbol{Z}_i, \boldsymbol{R}_i$ is defined as the product of univariate densities. Thus, from (2), the pdf of component k is also defined as

$$g_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = g_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) \prod_{j=1}^{d} p_{kj}^{r_{ij}}(x_{ij}) q_{kj}^{1-r_{ij}}(x_{ij}) \text{ with } g_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) = \prod_{j=1}^{d} \tau_{kj}^{r_{ij}} (1 - \tau_{kj})^{1-r_{ij}},$$

where $\boldsymbol{\tau}_k = (\tau_{k1}, \ldots, \tau_{kd})$, τ_{kj} is the probability that X_{ij} is observed given that subject i belongs to subpopulation k, $p_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 1$ and $q_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 0$. Thus, clustering is achieved by modeling, for each subpopulation, the marginal probability of missingness and the conditional density given that the variable is observed. Integrated out the unobserved variables $\boldsymbol{X}_i^{\text{miss}}$, we have

$$g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}), \text{ with } g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = g_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) \prod_{j=1}^{d} p_{kj}^{r_{ij}}(x_{ij}), \quad (3)$$

where $\boldsymbol{\theta}$ groups all the finite parameters (π_k and $\boldsymbol{\tau}_k$) and all the infinite parameters $p_{kj}(\cdot)$. Note, we do not need to estimate $q_{kj}(\cdot)$ for the clustering purpose but that this implies that we are not able to estimate the distribution of $\boldsymbol{X}_i \mid \boldsymbol{Z}_i$.

The following assumptions provide sufficient conditions for the model identifiability stated by Lemma 1 which is a consequence of Theorem 8 in Allman et al. (2009).

Assumption 1 *The p_{kj}'s are linearly independent, $\pi_k > 0$ and $\tau_{kj} > 0$.*

Lemma 1 *If Assumption 1 holds true, then the model defined by* (3) *is generically identifiable, up to label swapping.*

3 Maximizing of the smoothed likelihood

To perform parameter estimation, we extend the approach of Levine et al. (2011) that uses the smoothed likelihood to the case of mixed-type variables. Indeed, the observed variables contains continuous variables $\boldsymbol{x}_i^{\text{obs}}$ and binary variables $\boldsymbol{r}_i$. Note that the smoothing is only performed on the densities and thus on the distributions of $\boldsymbol{x}_i^{\text{obs}}$. Let S be the smoothing operator defined by $\mathcal{S}g_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i) = \prod_{j=1}^{d} (\mathcal{S}p_{kj}(x_{ij}))^{r_{ij}}$ where

$$\mathcal{S}p_{kj}(x_{ij}) = \int_{\Omega_j} \frac{1}{h} K\left(\frac{x_{ij} - u}{h}\right) p_{kj}(u) du,$$

where K is a kernel function and $h > 0$ its bandwidth. We consider the non linear smoothing operator defined by

$$\mathcal{N}g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = g_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) \exp\{\mathcal{S} \ln g_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i)\}. \text{ with } g_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i) = \prod_{j=1}^{d} p_{kj}^{r_{ij}}(x_{ij}).$$

The smoothed log-likelihood function is defined by

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) \right).$$

Parameter estimation is performed by maximizing the smoothed likelihood over $\boldsymbol{\theta}$. This maximization is achieved by a MM algorithm detailed in Du Roy de Chaumaray and Marbac (2020).

4 Numerical experiments

Tho illustrates the benefits of the proposed method, we compare, on simulated data, our proposed method to the following standard methods for clustering data with missingness: *GLMM* (Gaussian-Logit mixture model Miao et al. (2016)), *K-pod* (K-pod approach performed Chi et al. (2016)) and *NPimputed* (non parametric mixture on the imputed data where imputation is performed with the R package *missMDA* (Josse and Husson, 2016)). During all the experiments we use a Gaussian kernel with bandwidth $h = n^{-1/5}$.

We generate complete data from a bi-component mixture with unequal proportions ($\pi_1 = 1/3$ and $\pi_2 = 2/3$) and independence between variables within components such that $X_{ij} = \delta(Z_{i1} - Z_{i2}) + \varepsilon_{ij}$, where the ε_{ij} are independent from all the variables. Then, we add missing values from three scenarios: *MCA*: ($\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \boldsymbol{Z}_i) = (1 + \exp(\gamma))^{-1}$), *MNAR-1* ($\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \boldsymbol{Z}_i) = (1 + \exp(\gamma + z_{i1} - z_{i2}))^{-1}$), *MNAR-2* ($\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \boldsymbol{Z}_i) = (1 + \exp(\gamma + x_{ij}))^{-1}$). Thus, the parameters δ and γ allow to set the rates of misclassification error and missingness. We consider three distributions for ε_{ij}: standard Gaussian, Student with 3 degrees of freedom and Laplace.

We consider data sets composed by $n = 100$ observations and $d = 4$ variables. For each scenario, we generated 100 data sets. To compare the methods, we compute the Adjusted Rand index between the true partition and the estimators of the partition given by the methods. Results obtained for different rates of missingness and a theoretical missclassification rate of 5% are presented in Figure 1.

5 Conclusion

The proposed method allows continuous data set with nonignorable missingness to be clustered with no more assumption than the independence within components. Selecting the number of components is a difficult task that could be achieved by extending the approach of Kasahara and Shimotsu (2014) to the mixed-type data. Finally, a procedure of bandwidth selection should be investigated.

References

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

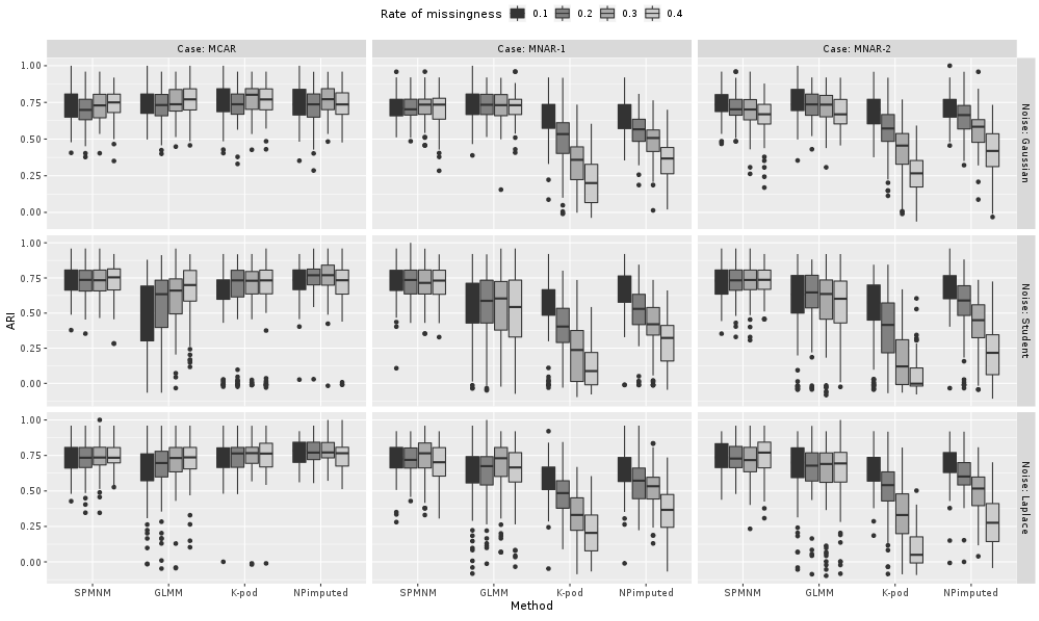


Figure 1: ARI obtained by the competing methods on 100 samples of 100 observations described by 4 variables for different rates of missingness and a theoretical rate of misclassification of 5%.

Chi, J. T., Chi, E. C., and Baraniuk, R. G. (2016). k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99.

Du Roy de Chaumaray, M. and Marbac, M. (2020). Clustering data with nonignorable missingness using semi-parametric mixture models. *arXiv preprint arXiv:2009.07662*.

Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

Kasahara, H. and Shimotsu, K. (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):97–111.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, pages 403–416.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Miao, W., Ding, P., and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683.

Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac [1], Mohammed Sedki [2], Christophe Biernacki [3], Vincent Vandewalle [4]

[1] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France. matthieu.marbac-lourdelle2@ensai.fr*

[2] *Univ. Paris-Sud and Inserm, France. mohammed.sedki@universite-paris-saclay.fr*

[3] *Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille. christophe.biernacki@inria.fr*

[4] *Univ. Lille, CHU Lille, ULR 2694 - METRICS and Inria, F-59000 Lille, France. vincent.vandewalle@univ-lille.fr*

Résumé. Nous étudions l'estimation des paramètres des modèles de régression avec des effets fixes par classe, lorsque la variable de classe est manquante alors que des variables liées à la classe sont disponibles. Ce problème peut être résolu en modélisant la distribution jointe de la variable cible et des variables liées à la classe. La stratégie habituelle d'estimation des paramètres pour ce modèle joint est une approche en deux étapes, commençant par l'apprentissage de la variable de la classe (étape de clustering) et ensuite l'insertion de son estimateur pour ajuster le modèle de régression (étape de régression). Toutefois, cette approche est sous-optimale car à la fois les estimateurs des paramètres de la régression sont biaisés et aussi elle n'utilise pas la variable cible pour le clustering. Ainsi, nous plaidons pour une approche d'estimation simultanée du clustering et de la régression, dans un cadre semi-paramétrique. Des expériences numériques illustrent les avantages de notre proposition en considérant différents modèles pour la distribution dans les classes et différents modèles de régression.

Mots-clés. clustering ; modèles de mélange ; modèles de régression ; modèles semi-paramétriques.

Abstract. We investigate the parameter estimation of regression models with fixed group effects, when the group variable is missing while group related variables are available. This problem can be solved by modeling the joint distribution of the target and of the group related variables. The usual parameter estimation strategy for this joint model is a two-step approach starting by learning the group variable (clustering step) and then plugging in its estimator for fitting the regression model (regression step). However, this approach is suboptimal since both regression estimates are biased and it does not make use of the target variable for clustering. Thus, we claim for a simultaneous estimation approach of both clustering and regression, in a semi-parametric framework. Numerical experiments illustrate the benefits of our proposition by considering wide ranges of distributions and regression models.

Keywords. clustering; finite mixture; regression model; semi-parametric model.

1 Introduction

The regression model with a fixed group effect considers that the intercept of the regression depends on the group from which the subject belongs (the intercept is common for subjects belonging to the same group but different for subjects belonging to different groups). However, in many applications, the group variable is not observed but other variables related to this variable are observed. For instance, suppose we want to investigate high blood pressure by considering the levels of physical activity among the covariates. In many cohorts, the level of physical activity of a subject is generally not directly available (because such a variable is not easily measurable) but many variables on the mean time spent doing different activities are available.

The estimation of a regression model with a fixed group effect is generally performed using a *two-step approach* as for instance in Epidemiology or in Economics. As a first step, a clustering on the individual based on the group related variables is performed to obtain an estimator of the group. As a second step, the regression model is fitted by using the estimator of the group variable among the covariates. However, since the group variable is estimated with error (class overlap), it is well-known that the resulting estimators of the parameters of regression are biased (Bertrand et al., 2017). The bias depends on the accuracy of the clustering step. Note that, although the target variable contains information about the group variable (and so is relevant for clustering), this information is not used in the two-step approach, leading to sub-optimal procedures.

We propose a new procedure (hereafter referred to as the *simultaneous approach*) that estimates simultaneously the clustering and the regression models in a semi-parametric frameworks (Hunter et al., 2011) thus circumventing the limits of the standard procedure (biased estimators). We demonstrate that this procedure improves both the estimators of the partition and regression parameters. We focus on semi-parametric mixture where the component densities are defined as a product of univariate densities (Chauveau et al., 2015), which is identifiable if the univariate densities are linearly independent and if at least three variables are used for clustering (Allman et al., 2009). Semi-parametric inference is achieved by a maximum smoothed likelihood approach (Levine et al., 2011) via a Maximization-Minimization (MM) algorithm (Hunter and Lange, 2004).

The presentation is organized as follows. Section 2 introduces a general context where a statistical analysis requires both methods of clustering and prediction, and it presents the standard approach that estimates the parameters in two steps. Section 3 shows that a procedure that allows a simultaneous estimation of the clustering and of the regression parameters generally outperforms the two-step approach. Section 4 discusses about numerical experiments, which are not given in this long summary but will be presented during the talk. More details about the work presented can be found in Marbac et al. (2020).

2 Embedding clustering and prediction models

2.1 Data presentation

Let $(V^\top, X^\top, Y)^\top$ be the set of the random variables where $V = (U^\top, Z^\top)^\top$ is a $d_V = d_U + K$ dimensional vector used as covariates for the prediction of the univariate variable $Y \in \mathbb{R}$, X is a d_X dimensional vector and $Z = (Z_1, \ldots, Z_K)^\top \in \mathcal{Z}$ is a categorical variable with K levels. The variable Z indicates the group membership such that $Z_k = 1$ if the subject belongs to cluster k and otherwise $Z_k = 0$. The realizations of $(U^\top, X^\top, Y)^\top$ are observed but the realizations of Z are unobserved. Thus, X is a set of proxy variables used to estimate the realizations of Z. Considering the high blood pressure example, Y corresponds to the diastolic blood pressure, U is the set of observed covariates (gender, age, alcohol consumption, obesity and sleep quality), X is the set of covariates measuring the level of physical activity and Z indicates the membership of a group of subjects with similar physical activity behaviours. The observed data are n independent copies of $(U^\top, X^\top, Y)^\top$ denoted by $\mathbb{U} = (u_1, \ldots, u_n)^\top$, $\mathbb{X} = (x_1, \ldots, x_n)^\top$ and $\mathbb{Y} = (y_1, \ldots, y_n)^\top$ respectively. The n unobserved realizations of Z are denoted by $\mathbb{Z} = (z_1, \ldots, z_n)^\top$.

2.2 Introducing the joint predictive clustering model

Regression model Let a loss function be $\mathcal{L}(\cdot)$ and $\rho(\cdot)$ its piecewise derivative. The loss function $\mathcal{L}$ allows the regression model of Y on V to be specified with a fixed group effect given by

$$Y = V^\top \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0, \tag{1}$$

where $\beta = (\gamma^\top, \delta^\top)^\top \in \mathbb{R}^{d_V}$, $\gamma \in \mathbb{R}^{d_U}$ are the coefficients of U, $\delta = (\delta_1, \ldots, \delta_K)^\top \in \mathbb{R}^K$ are the coefficients of Z (*i.e.,* the parameters of the group effect), and ε is the noise. Note that for reasons of identifiability, the model does not have an intercept. The choice of $\mathcal{L}$ allows many models to be considered and, among them, one can cite the mean regression (with $\mathcal{L}(t) = t^2$ and $\rho(t) = 2t$), the τ-quantile regression (with $\mathcal{L}(t) = |t| + (2\tau - 1)t$ and $\rho(\varepsilon) = \tau - \mathbb{1}_{\{\varepsilon \leq 0\}}$), the τ-expectile regression (with $\mathcal{L}(t) = |\tau - \mathbf{1}\{t \leq 0\}|t^2$ and $\rho(t) = 2t((1-\tau)\mathbf{1}\{t \leq 0\} + \tau \mathbf{1}\{t > 0\})$.

The restriction on the conditional moment of $\rho(\varepsilon)$ given V is sufficient to define a model and allows for parameter estimation. However, obtaining maximum likelihood estimate (MLE) needs specific assumptions on the noise distribution. For instance, parameters of the mean regression can be consistently estimated with MLE by assuming a centred Gaussian noise. Similarly, the parameters of τ-quantile (or τ-expectile) regression can be consistently estimated with MLE by assuming that the noise follows an asymmetric Laplace (or an asymmetric normal) distribution. Hereafter, we denote the density of the noise ε by f_ε.

Clustering model The distribution of X given $Z_k = 1$ is defined by the density $f_k(\cdot)$. Therefore, the marginal distribution of X is a mixture model defined by the density

$$f(x;\vartheta) = \sum_{k=1}^{K} \pi_k f_k(x) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j), \tag{2}$$

where $\vartheta = \{\pi_k, f_k; k = 1, \ldots, K\}$, $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$ and where f_k is the density of component k defined as the product of univariate densities f_{kj} in the semi-parametric setting.

Joint clustering and regression model The joint model assumes that Z explains the dependency between Y and X (*i.e.,* Y and X are conditionally independent given Z) and that U and $(X^\top, Z^\top)$ are independent. Moreover, the distribution of (X, Y) given U is also a mixture model defined by the density (noting $\theta = \{\vartheta\} \cup \{\delta_k; k = 1, \ldots, K\} \cup \{\gamma, f_\varepsilon\}$)

$$f(x, y|u;\theta) = \sum_{k=1}^{K} \pi_k f_k(x) f_\varepsilon(y - u^\top\gamma - \delta_k), \tag{3}$$

where, for $k = 1, \ldots, K$ we have

$$\mathbb{E}[\rho(Y - U^\top\gamma - \delta_k)|U, Z_k = 1] = 0. \tag{4}$$

Moment condition The following lemma gives the moment equation verified on the joint model. It will be used later to justify the need for a simultaneous approach. It shows an equivalence between the moment equation which permits to understand why the two-step approach is biased and which justifies the use of the unified procedure.

Lemma 1. *Let an identifiable model defined by* (3) *and* (4)*, for any x and k. Then, noting* $r_k^{X,Y}(x, y) = \frac{\pi_k f_k(x) f_\varepsilon(y - u^\top\gamma - \delta_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x) f_\varepsilon(y - u^\top\gamma - \delta_\ell)}$*, $\beta = (\delta^\top, \gamma^\top)^\top$ is the single parameter satisfying*

$$\forall k = 1, \ldots, K,\ \mathbb{E}[r_k^{X,Y}(X, Y)\rho(Y - u^\top\gamma - \delta_k)|U, X] = 0. \tag{5}$$

3 The proposed simultaneous estimation procedure

Based on Lemma 1, we have shown in Marbac et al. (2020) that performing a two-step approach, thus performing the regression based on the clustering step output, provides a suboptimal classification rule because the classification neglects the information given by Y. Consequently, we circumvent this issue by using a simultaneous semi-parametric approach, also avoiding bias of parametric miss-specified models.

Semi-parametric model In this section, we consider the semi-parametric version of the model defined by (3) where the densities of the components are assumed to be a product of univariate densities. Thus, we have

$$f(y, x \mid u; \theta) = \sum_{k=1}^{K} \pi_k f_k(y, x \mid u; \theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j) f_\varepsilon(y - u^\top \gamma - \delta_k), \tag{6}$$

where θ groups all the finite and infinite parameters and β is such that (5) holds. A sufficient condition implying identifiability for model (6) is that the marginal distribution of X is identifiable and thus a sufficient condition is to consider linearly independent densities f_{kj}'s and $d_X \geq 3$ (Allman et al., 2009). For sake of simplicity we will note $w = (x^\top, y)^\top$ with $w \in \mathbb{R}^{d_X+1}$, such that $f(y, x \mid u; \theta) = \sum_{k=1}^{K} \pi_k f_k(w \mid u; \theta)$.

Majorization-Minorization algorithm Parameter estimation is achieved via a Majorization-Minorization algorithm. Given an initial value $\theta^{[0]}$, this algorithm iterates between a majorization and a minorization steps. Thus, an iteration $[r]$ is defined by

- Majorization step: $t_{ik}^{[r-1]} \propto \pi_k^{[r-1]} \left(\mathcal{N} f_k^{[r-1]}\right)(w_i \mid u_i; \theta^{[r-1]})$, with $\mathcal{N} f_k$ the exponential of the smoothed log-density in class k (Levine et al., 2011).
- Minorization step:
  1. Updating the parametric elements

$$\pi_k^{[r]} = \frac{1}{n} \sum_i t_{ik}^{[r-1]} \text{ and } \beta^{[r]} = \arg\min_\beta \sum_{i,k} t_{ik}^{[r-1]} \rho(y_i - u_i^\top \gamma - \delta_k).$$

  2. Updating the nonparametric elements

$$f_{kj}(a) = \frac{1}{n\pi_k^{[r]}} \sum_i t_{ik}^{[r-1]} K_h(x_{ij} - a) \text{ and } f_\varepsilon(a) = \frac{1}{n} \sum_{i,k} t_{ik}^{[r-1]} K_h(y_i - u_i^\top \gamma^{[r]} - \delta_k^{[r]} - a),$$

     with K_h the rescale kernel function of bandwith h considered in the smoothing.

The Majorization-Minorization algorithm is monotonic for the smoothed log-likelihood. It is a direct consequence of the monotony of the algorithm of Levine et al. (2011) where we use the fact that, in order to satisfy the moment condition defined in (5) of Lemma 1, we must have $\beta^{[r]} = \arg\min_\beta \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^{[r-1]} \rho(y_i - u_i^\top \gamma - \delta_k)$.

4 Numerical experiments and conclusion

In experiments presented in Marbac et al. (2020), we have considered several types of regressions in a semi-parametric framework. In a first time we have compared the simultaneous and the two-step approaches in a parametric and semi-parametric framework. In a second time we have shown that robust regressions can be easily used with the semi-parametric approach and improve the estimators of the regression parameters. In a third time, we have shown that the semi-parametric method permits to consider asymetric losses (quantile or expectile regressions). An application the high blood pressure has also been studied where we have considered simultaneously the clustering of subjects based on their physical activity and the use of this variable in a regression model on the diastolic blood pressure.

The main conclusion is that simultaneously performing the clustering and the estimation of the regression model improves the accuracy of both the partition and of the regression parameters. The approach can be applied to a wide range of regression models, and avoids bias in the estimation compared with parametric models.

References

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Bertrand, A., Legrand, C., Léonard, D., and Van Keilegom, I. (2017). Robustness of estimation methods in a survival cure model with mismeasured covariates. *Computational Statistics & Data Analysis*, 113:3–18.

Chauveau, D., Hunter, D. R., Levine, M., et al. (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys*, 9:1–31.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Hunter, D. R., Richards, D. S. P., and Rosenberger, J. L. (2011). *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger, the Pennsylvania State University, USA, 23-24 May 2008*. World Scientific.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, pages 403–416.

Marbac, M., Sedki, M., Biernacki, C., and Vandewalle, V. (2020). Simultaneous semi-parametric estimation of clustering and regression. preprint, `https://hal.inria.fr/hal-03090573`.

Co-clustering of evolving count matrices in pharmacovigilance with the dynamic latent block model

Giulia Marchello [1], Audrey Fresse[2], Marco Corneli[3] & Charles Bouveyron [1]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.*

[2] *Université Côte d'Azur, Department of Clinical Pharmacology, Pasteur Hospital, Nice, France.*

[3] *Université Côte d'Azur, Inria, Maison de la Modélisation des Simulations et des Interactions (MSI), Maasai team, Nice, France.*

E-mail: giulia.marchello@inria.fr

Résumé. L'objectif de ce travail est d'analyser les notifications d'effets indésirables (EIs) recueillies par le Centre Régional de Pharmacovigilance de Nice (France) entre 2010 et 2020. La détection actuelle des signaux par les experts est malheureusement incomplète, donc nous étudions ici une méthode automatisée de détection des signaux. À cette fin, nous introduisons un modèle de co-clustering génératif, nommé modèle de bloc latent dynamique (dLBM), qui étend le modèle de bloc latent classique au cas des données de comptage dynamique. Le temps continu est géré en partitionnant la période de temps considérée, permettant la détection de coupures temporelles dans les signaux. Un algorithme SEM-Gibbs est proposé pour l'inférence et le critère ICL est utilisé pour la sélection du modèle. L'application à l'ensemble de données ADR a souligné que le dLBM était non seulement capable d'identifier des clusters cohérentes avec les connaissances rétrospectives, mais aussi de détecter des comportements atypiques, dont les professionnels de santé n'étaient pas conscients.

Mots-clés. Co-clustering, pharmacovigilance, modèle de bloc latent, matrices de comptage dynamique, algorithme SEM-Gibbs.

Abstract. The purpose of this work is to analyze the notifications of adverse drug reactions (ADR) gathered by the Regional Center of Pharmacovigilance of Nice (France) between 2010 to 2020. As the current expert detection of safety signals is unfortunately incomplete, we investigate here an automatized method of safety signal detection. To this end, we introduce a generative co-clustering model, named dynamic latent block model (dLBM), which extends the classical binary latent block model to the case of dynamic count data. The continuous time is handled by partitioning the considered time period, allowing the detection of temporal breaks in the signals. A SEM-Gibbs algorithm is proposed for inference and the ICL criterion is used for model selection. The application to ADR dataset pointed out that dLBM was not only able to identify clusters that are coherent with retrospective knowledge, but also to detect atypical behaviors, which the health professionals were unaware.

Keywords. Co-clustering, pharmacovigilance, latent block model, dynamic count matrices, SEM-Gibbs algorithm.

1 Introduction

One of the missions of the Regional Centers of Pharmacovigilance (RCPVs) is safety signal detection. However, the method currently used, i.e. manual expert detection of safety signals by the RCPV, despite being unavoidable, has the disadvantage of being incomplete due to its workload. This is why, developing automatized method of safety signal detection is currently a major issue in pharmacovigilance. In such a context, clustering may play an important role in summarizing the information carried out by pharmacovigilance data and identifying patterns of interest. It would be indeed of interest to both cluster the drugs and the adverse reactions to help medical experts in their tasks.

2 The dynamic latent block model (dLBM)

The main goal of this model is the simultaneous clustering of rows and columns of high-dimensional sparse matrices in a dynamic time framework. The data we consider are organized such that the rows (drugs in pharmacovigilance application) are indexed by $i = 1, \dots N$ and the columns (adversarial effects) by $j = 1, \dots, P$. Moreover, we consider a fixed time period $[0, T]$ during which the total number of rows, N, and columns, P, is fixed. We indicate as $\mathcal{X}(t)$ the $N \times P$ matrix that represents the cumulative number of interactions between i and j at time $t \in [0, T]$. According to the latent block model (Govaert and Nadif, 2010), rows and columns of $\mathcal{X}(t)$ are assumed to be clustered respectively into K and L groups, such that the data belonging to the same block are independent and identically distributed. More formally, the latent structure of $\mathcal{X}(t)$ is identified by:

- $Z := \{z_{ik}\}_{i\in 1,\dots,N, k\in 1,\dots,K}$ represents the clustering of rows into K groups: $\mathcal{A}_1, ..., \mathcal{A}_K$. The row i belongs to cluster $\mathcal{A}_k$ iff $z_{ik} = 1$;
- $W := \{w_{j\ell}\}_{j\in 1,\dots,P, \ell\in 1,\dots,L}$ represents the clustering of columns into L groups: $\mathcal{B}_1, ..., \mathcal{B}_L$. The column j belongs to cluster $\mathcal{B}_\ell$ iff $w_{j\ell} = 1$.

Moreover, Z and W are assumed to be independent and distributed according to multinomial distributions:

$$p(Z|\gamma) = \prod_{k=1}^{K} \gamma_k^{|\mathcal{A}_k|}, \qquad p(W|\rho) = \prod_{\ell=1}^{L} \rho_\ell^{|\mathcal{B}_\ell|},$$

where $\gamma_k = \mathbb{P}\{z_{ik} = 1\}$, $\rho_\ell = \mathbb{P}\{w_{j\ell} = 1\}$, $\sum_{k=1}^{K} \gamma_k = 1$, $\sum_{\ell=1}^{L} \rho_\ell = 1$, and $|\mathcal{A}_k|$ and $|\mathcal{B}_\ell|$ respectively represent the number of rows in cluster $\mathcal{A}_k$ and the number of columns in cluster $\mathcal{B}_\ell$.

Modeling the dynamic framework A possible approach for the dynamic modeling relies on non-homogeneous Poisson processes (NHPPs), thus assuming that $\{\mathcal{X}_{ij}(.)\}_{i,j}$ are independent point processes, with instantaneous intensity functions λ. We further assume that the intensity function only depends on the respective clusters of row i and column j:

$$\mathcal{X}_{ij}(t) \mid z_{ik}, w_{j\ell} = 1 \sim \mathcal{P}\left(\int_0^t \lambda_{k\ell}(u)du\right). \tag{1}$$

In order to ease the understanding of the dynamic model and to make the inference tractable, we also operate a clustering over the time dimension. Let us first introduce a discretization of the considered time interval $[0, T]$, as follows:

$$0 = t_0 < t_1 < \cdots < t_U = T, \tag{2}$$

where the U intervals, $I_u = [t_{u-1}, t_u[$, will also be clustered. The number of interactions between i and j on the time interval I_u can be therefore summarized by:

$$X_{iju} := \mathcal{X}_{ij}(t_u) - \mathcal{X}_{ij}(t_{u-1}), \qquad \forall (i, j, u), \tag{3}$$

where $\mathcal{X}_{ij}(t_u)$ represents the cumulative number of interactions at time t_u between i and j. Since our goal is to perform clustering over the time dimension as well, each time interval $I_1, \ldots, I_U$ is also assumed to be assigned to a hidden time cluster $\mathcal{D}_1, \ldots, \mathcal{D}_c$. A specific time cluster can occur more than once in the temporal line when a similar interactivity pattern is repeated in time. To model the membership to time clusters, a new latent variable S has to be introduced. As for Z and W, we assume that S follows a multinomial distribution:

$$p(S \mid \delta) = \prod_{c=1}^{C} \delta_c^{|\mathcal{D}_c|}, \tag{4}$$

where $\delta_c = \mathbb{P}\{s_{uc} = 1\}$; $\sum_{c=1}^{C} \delta_c = 1$ and $\mid \mathcal{D}_c \mid$ represents the number of time intervals in the cluster $\mathcal{D}_c$. Once these additional assumptions have been made, we can write:

$$X_{iju} | z_{ik} w_{j\ell} s_{uc} = 1 \sim \mathcal{P}(\lambda_{k\ell c} \Delta_u), \tag{5}$$

where Δ_u indicates the length of the interval I_u.We assume that Δ_u is constant, $\Delta_u = \Delta$. We can finally set $\Delta = 1$ without loss of generality. Thus, it holds that:

$$p(X_{iju} \mid z_{ik} w_{j\ell} s_{uc} = 1, \lambda_{k\ell c}) = \left(\frac{(\lambda_{k\ell c})^{X_{iju}}}{X_{iju}!} \exp\left(-\lambda_{k\ell c}\right)\right). \tag{6}$$

It is now possible to write the complete data likelihood of the model:

$$p(X, Z, W, S|\gamma, \rho, \delta, \lambda) = p(Z|\gamma)p(W|\rho)p(S|\delta)p(X|Z, W, S, \lambda), \tag{7}$$

where $p(Z|\gamma)$, $p(W|\rho)$ and $p(S|\delta)$ were defined in the previous section. The conditional distribution of X, given Z, W, and S, can be easily obtained from Eq. (6) by independence:

$$p(X|Z, W, S, \lambda) = \prod_{k,\ell,c} \left(\frac{(\lambda_{k\ell c})^{R_{k\ell c}}}{P_{k\ell c}} \exp\left(- \mid \mathcal{A}_k \mid\mid \mathcal{B}_\ell \mid\mid \mathcal{D}_c \mid \lambda_{k\ell c}\right) \right), \tag{8}$$

where $R_{k\ell c} = \sum_{i=1}^{N} \sum_{j=1}^{P} \sum_{u=1}^{U} z_{ik} w_{j\ell} s_{uc} X_{iju}$ and $P_{k\ell c} = \prod_{i=1}^{N} \prod_{j=1}^{P} \prod_{u=1}^{U} (z_{ik} w_{j\ell} s_{uc} X_{iju})!$.

Model inference We look for a way to maximize the log-likelihood in order to obtain the estimation of the model parameters, θ. In the co-clustering case, the EM algorithm is computationally unfeasible, the idea is to use a stochastic version of it, known as SEM-Gibbs, proposed by Keribin et al. (2010). Thanks to the Gibbs sampler within the SE step a partition for Z, W and S is generated without computing the joint distribution. The algorithm starts with initial values for the parameter set $\theta^{(0)}$, the column clusters $W^{(0)}$ and the time clusters $S^{(0)}$. Regarding the burn-in period, after a certain number of iterations of the algorithm, we can obtain the final parameters estimation by computing the mean of the sampled distribution. The optimal values for Z, W and S are estimated by the mode of their sample distributions.

Model selection Up to now, we have assumed that the number of row clusters (K), column clusters (L) and time clusters (C) was known. However, for real data sets, this assumption is of course unrealistic. For this reason, our purpose in this section is to define a model selection criterion that can automatically identify the optimal number of clusters. We rely on ICL (Integrated Completed Likelihood, Biernacki et al. (2000)):

$$\begin{aligned} ICL(K, L, C) = \log p(X, \hat{Z}, \hat{W}, \hat{S}; \hat{\theta}) - \frac{K-1}{2} \log N + \\ - \frac{L-1}{2} \log P - \frac{C-1}{2} \log U - \frac{KLC}{2} \log(NPU) \end{aligned} \tag{9}$$

The triplet $(\hat{K}, \hat{L}, \hat{C})$ that leads to the highest value for the ICL is considered as the most meaningful for those data.

3 Analysis of the adverse drug reaction dataset

This section considers a large dataset consisting of ADR data collected by the Regional Center of Pharmacovigilance (RCPV), located in the University Hospital of Nice (France).

The center covers an area of over 2.3 million inhabitants. A time horizon of 10 years is considered, from January 1^{st}, 2010 to September 30^{th}, 2020, the unity measure for time intervals is a month ($\Delta_u = \Delta = 1$ month). The overall dataset is made of by 44,269 declarations. We only considered drugs and ADRs that were notified more than 10 times over the 10 years. During this period, an extremely uncommon behavior happened in the progress of notifications to the RCPV. In fact, in 2017 an unexpected rise of reports for ADRs happened concerning a specific drug called Lévothyrox®. This has been marketed in France for about 40 years as a treatment for hypothyroidism and, in 2017, a new formula was introduced on the market. The Lévothyrox® case had an extremely high media coverage in France: Lévothyrox® spontaneous reports represent almost the 90% of all the spontaneous notifications that the RCPV received in 2017 (Viard et al., 2019). Behind those very visible effects, many ADR signals need to be detected for obvious public health reasons. In particular, those data also contain ADR reports regarding Médiator®, which is here far less visible, but also led to many avoidable serious cardiovascular diseases. This is why, we expect dLBM to be a useful tool to reveal such hidden signals.

3.1 Summary of the results

We have run dLBM for different values of K, L and C, we tested row (here drugs), column (here ADRs) and time groups ranging from 1 to 12. The ICL criterion identified the optimal values as: $\hat{K} = 7$, $\hat{L} = 10$, $\hat{C} = 6$. Figure 1 shows the frequency of the declarations received by the RCPV from 2010 to 2020, sorted by month, where the colors represent the identified time clusters. Figure 2 shows the evolution of the relationship between drug clusters and ADR clusters over time. In fact, each panel represents a cluster of drugs and within them each line identifies a cluster of ADRs and its intensity changes over time. In this application to pharmacovigilance, dLBM proved to be a very useful tool for identifying phenomena that would have been difficult to detect otherwise, even by an expert eye. In fact, dLBM revealed that in addition to Lévothyrox® health crisis, which was the one with the widest media coverage, two other major events have occurred. The first one concerning Médiator®, which took place in 2009-2010, and the second one concerning Mirena®, which took place in the first half of 2017. In addition, dLBM was also able to put in light some unexpected variations of notifications such as an under-notification of bleeding related ADRs during Lévothyrox® crisis. Another thing that dLBM has highlighted is the existence of 3 different phases during the Lévothyrox® crisis corresponding to the reporting peak, the marketing period

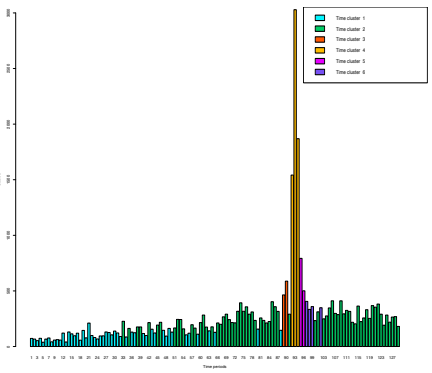


Figure 1: Reports received by the RCPV, colors represent the time clusters.

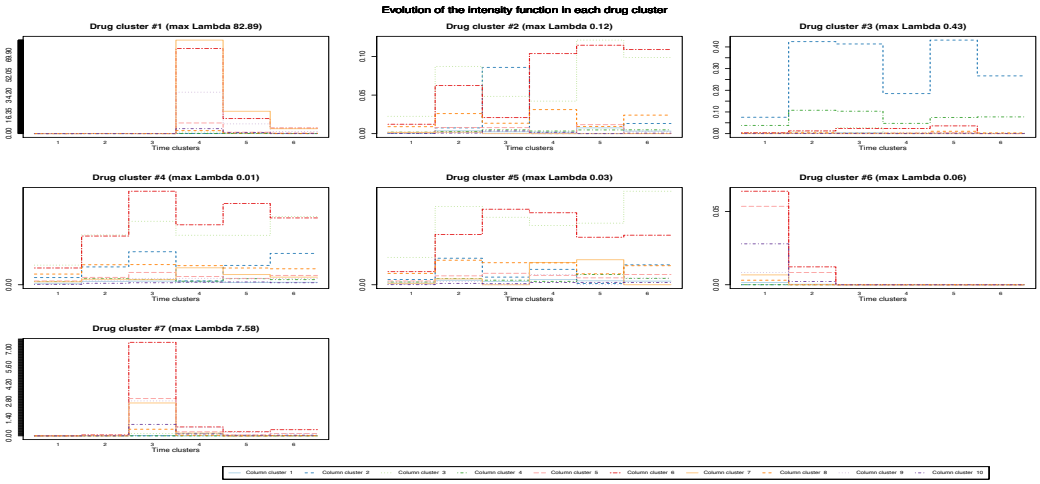


Figure 2: Evolution of the relation between each drug cluster and the all ADR clusters over time. Each color corresponds to a different cluster of adverse drug reaction.

of generics and the end of the crisis, respectively. Those phases were not noticed by the RCPV staff during the Lévothyrox® crisis. In general, we can conclude that dLBM could be extremely useful as a routine tool for signal detection, since it might help health professionals to identify structural changes or patterns of interest and, perhaps, prevent some of the consequences a health crisis can lead to.

References

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425, 2010.

C. Keribin, G. Govaert, and G. Celeux. Estimation d'un modèle à blocs latents par l'algorithme sem. 2010.

D. Viard, N. Parassol-Girard, S. Romani, E. Van Obberghen, F. Rocher, S. Berriri, and M.-D. Drici. Spontaneous adverse event notifications by patients subsequent to the marketing of a new formulation of levothyrox® amidst a drug media crisis: atypical profile as compared with other drugs. *Fundamental & clinical pharmacology*, 33(4):463–470, 2019.

NOUVEL ALGORITHME STATISTIQUE DE COMPARAISON DE DEUX ECHANTILLONS INDEPENDANTS DANS LE CAS D'UNE VARIABLE ORDINALE : APPLICATION AUX DOMAINES DE LA SANTE

Abdelghafour Marfak[1] & Ibtissam Youlyouz-Marfak[2]

[1]*École Nationale de Santé Publique, Rabat, Maroc*
[2]*Institut Supérieur des Sciences de la Santé, Université Hassan Premier de Settat, Maroc*

Résumé

Pour comparer deux moyennes, estimées sur deux échantillons indépendants (X et Y), habituellement le test de Student est employé dans le cas de la normalité des distributions. Dans le cas où la variable est de type ordinal, le test de Mann-Whitney est essentiellement appliqué. Il consiste à mettre ensemble les réponses des 2 groupes X et Y puis à les classer. L'inférence statistique est basée ensuite sur la somme des rangs des sujets des deux échantillons X et Y. Dans le cas de mise en évidence d'une différence entre X et Y, le test de Mann-Whitney ne permet pas de quantifier cette différence statistique. Récemment, nous avons proposé un nouvel algorithme basé sur la méthode RIDIT (Relative to an Identified Distribution) que nous l'avons nommé « Improved RIDIT » (Marfak et al. 2020). A la différence du test de Mann-Whitney, Improved RIDIT apporte de nouvelles statistiques dans le cas d'une variable ordinale telles que l'Augmentation du Risque Absolu (ARI : Absolute Risk Increase)/Reduction du Risque Absolu (ARR : Absolute Risk Reduction), l'odds ratio ordinal ($odds_{Ordinal}$) et le Nombre de sujet Nécessaire à Traiter (NNT). Dans cet article, nous présentons la méthode Improved RIDIT et un exemple d'application dans le domaine de la santé.

Mots-clés. Tests Statistiques, Improved RIDIT, odds Ordinal, Risque Absolu, Nombre de Sujet Traités.

Abstract

To compare two means, estimated on two independent samples (X and Y), usually the Student's test is used in the case of the normality of the distributions. In the case of ordinal variables, the Mann-Whitney test is essentially applied. It consists in putting together the answers of the 2 groups X and Y and classifying them. Therefore, statistical inference is based on the sum of the ranks of the subjects of the two samples X and Y. When a difference between X and Y is observed, the Mann-Whitney is not able to quantify this difference. statistical. Recently, we proposed a new algorithm based on the RIDIT (Relative to an Identified Distribution) method that we named "Improved RIDIT" (Marfak et al. 2020). Unlike the Mann-Whitney test, Improved RIDIT provides new statistics in the case of an ordinal variable such as Absolute Risk Reduction (ARR)/Absolute Risk Increase (ARI), the ordinal odds ratio ($odds_{Ordinal}$) and the Number of subjects Needed to Treat (NNT). In this article, we present the Improved RIDIT method and an example of application for health-related quality of life.

Keywords. Statistical tests, Improved RIDIT, odds Ordinal, Absolute Risk, Number Needed to Treat.

1. Introduction

Le test de Mann-Whitney est essentiellement appliqué pour comparer deux échantillons indépendants dans le cas d'une variables ordinale. Il consiste à mettre ensemble les réponses des 2 groupes X et Y puis à les classer. L'inférence statistique est basée ensuite sur la somme des rangs des sujets des deux échantillons X et Y. Dans le cas de mise en évidence d'une différence entre X et Y, le test de Mann-Whitney ne permet pas de quantifier cette différence statistique. Récemment, nous avons proposé un nouvel algorithme basé sur la méthode RIDIT (Relative to an Identified Distribution) que nous l'avons nommé « Improved RIDIT » (Marfak et al. 2020). A la différence du test de Mann-Whitney, la méthode Improved RIDIT apporte de nouvelles statistiques dans le cas d'une variable ordinale telles que l'Augmentation du Risque Absolu (ARI : Absolute Risk Increase)/Reduction du Risque Absolu (ARR : Absolute Risk Reduction), l'odds ratio ordinal ($odds_{Ordinal}$) et le Nombre de sujet Nécessaire à Traiter (NNT). Dans cet article, nous présentons la méthode Improved RIDIT et un exemple d'application dans le domaine de la santé.

Dans cet article, nous présentons dans une première partie les fondements de la méthode Improved RIDIT et l'algorithme de calcul. Dans une deuxième partie nous appliquons la méthode à des données de la qualité de vie.

2. La méthode Improved RIDIT

Considérons deux échantillons indépendants **Y** (témoins ou recevant un traitement dans le cadre d'un essai clinique) et **X** (patients atteints d'une maladie chronique ou recevant un traitement placebo). Supposons qu'une variable à l'étude (exemple, la douleur) est évaluée par chacun des sujets des deux échantillons sur une échelle de Likert à **k** niveaux. Les données obtenues par l'ensemble des sujets sont résumées dans un tableau de contingence où les lignes se réfèrent aux échantillons **Y** et **X** et les colonnes se réfèrent aux niveaux de l'échelle de mesure (**Tableau 1**). La variable ordinale est classée du niveau le moins grave (exemple, pas de douleur est codé «1») au niveau le plus grave (exemple, douleur extrême est codé «5»).

Tableau 1 : format des données pour réaliser les calculs par la méthode Improved RIDIT

Level of severity (*l*)	**1**	**…**	**2**	**k**	**Total**
Echantillon Y	$f_Y^{(1)}$	…	$f_Y^{(k-1)}$	$f_Y^{(k)}$	$\mathbf{n_y}$
Echantillon X	$f_X^{(1)}$	…	$f_X^{(k-1)}$	$f_X^{(k)}$	$\mathbf{n_x}$
Total	t_1	…	t_{k-1}	t_k	$\mathbf{N = n_y + n_x}$

k est le nombre de niveaux de l'échelle Likert. Pour chaque niveau, $l = 1, \dots, k$, $f_Y^{(l)}$ and $f_X^{(l)}$ dénote les fréquences observées respectivement chez les sujets the des échantillons Y and X.

Notons y_j (j = 1,…, n_y) les n_y observations de l'échantillon **Y** et x_i (i = 1,…, n_x) les n_x observations de l'échantillon **X**. Soit la fonction **ω** définie par :

$$\omega_{ij} = \begin{cases} +1 & if\ x_i > y_j \\ 0 & if\ x_i = y_j \\ -1 & if\ x_i < y_j \end{cases} \quad (eq.1)$$

A partir de la fonction **ω**, nous définissons les probabilités :

$$\pi^+ = P[X > Y]\ (eq.2)$$

$$\pi^0 = P[X = Y]\ (eq.3)$$

$$\pi^- = P[X < Y] \ (eq. 4)$$

π^+, π^0 and π^- indiquent respectivement que des sujets de l'échantillon **X** sont dans un niveau plus sévère, similaire ou moins sévère que des sujets de l'échantillon **Y**. Le test Improved RIDIT est basé sur la statistique :

$$W = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \omega_{ij} \ (eq. 5)$$

Avec $$E[W] = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} E\left[\omega_{ij}\right] \ (eq. 6)$$

A partir des équations eq. 1 et eqs. 2-4, nous obtenons :

$$E[W] = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (\pi^+ - \pi^-) \ (eq. 7)$$

Alors, $$E[W] = n_y n_x (\pi^+ - \pi^-) \ (eq. 8)$$

Une estimation de $(\pi^+ - \pi^-)$ est :

$$\hat{\pi}^+ - \hat{\pi}^- = \frac{W}{n_y n_x} \ (eq. 9)$$

A partir de (eq.1) et le **Tableau 1**, nous réécrivons (eq.5) :

$$W = \sum_{l=1}^{k-1} \left(f_Y^{(l)} \sum_{p=l+1}^{k} f_X^{(p)} \right) - \sum_{l=2}^{k} \left(f_Y^{(l)} \sum_{p=1}^{l-1} f_X^{(p)} \right) \ (eq. 10)$$

La variance de W est :

$$Var(W) = \frac{n_y n_x (N+1)}{3} \left(1 - \frac{\sum_{l=1}^{k} (t_l^3 - t_l)}{N^3 - N} \right) \ (eq. 11)$$

A partir des équations eq.3, eq.9 et la propriété de la probabilité $\pi^0 + \pi^- + \pi^+ = 1$, les estimations des ex-aequo π^0 et π^+ et π^- sont :

$$\hat{\pi}^0 = \frac{1}{n_y n_x} \sum_{l=1}^{k} f_Y^{(l)} f_X^{(l)} \ (eq. 12)$$

$$\hat{\pi}^+ = \frac{1}{2} \left(1 - \hat{\pi}^0 + \frac{W}{n_y n_x} \right) \ (eq. 13)$$

$$\hat{\pi}^- = \frac{1}{2} \left(1 - \hat{\pi}^0 - \frac{W}{n_y n_x} \right) \ (eq. 14)$$

L'estimation de π^- and π^+ exige le calcul de W (eq.10), ce qui n'est pas évident. Nous avons proposé un simple algorithme (Marfak et al. 2020) :

1. Transformmer la table de contingence (Table 1) en matrice carrée $\boldsymbol{\omega}_{(\boldsymbol{k},\boldsymbol{k})}$ dont les élements sont les produits des fréquences $f_X^{(l)}$ and $f_Y^{(m)}$ ($l, m = 1,\ldots, k$), avec k est le niveau de la variable ordinale :

$$\boldsymbol{\omega} = \begin{pmatrix} f_Y^{(1)} f_X^{(1)} & f_Y^{(1)} f_X^{(2)} & \cdots & f_Y^{(1)} f_X^{(k-1)} & f_Y^{(1)} f_X^{(k)} \\ f_Y^{(2)} f_X^{(1)} & f_Y^{(2)} f_X^{(2)} & \cdots & f_Y^{(2)} f_X^{(k-1)} & f_Y^{(2)} f_X^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_Y^{(k-1)} f_X^{(1)} & f_Y^{(k-1)} f_X^{(2)} & \cdots & f_Y^{(k-1)} f_X^{(k-1)} & f_Y^{(k-1)} f_X^{(k)} \\ f_Y^{(k)} f_X^{(1)} & f_Y^{(k)} f_X^{(2)} & \cdots & f_Y^{(k)} f_X^{(k-1)} & f_Y^{(k)} f_X^{(k)} \end{pmatrix}$$

   Regardant les éléments de la matrice $\boldsymbol{\omega}_{(\boldsymbol{k},\boldsymbol{k})}$, on peut noter :
   - La somme de tous les éléments $\boldsymbol{\omega}_{(\boldsymbol{k},\boldsymbol{k})}$ est égale à $n_x n_y$.
   - La somme des éléments au-dessus de la diagonale est équivalent au terme gauche de (eq. 10). Ces valeurs $f_Y^{(m)} f_X^{(l)}$ $(l > m)$ indiquent qu'un sujet tiré au sort de l'échantillon **X** est dans un niveau élevé qu'un sujet tiré au sort de l'échantillon **Y**.

- La somme des éléments au-dessous de la diagonale est équivalent au terme droite de (eq. 10). Ces valeurs $f_Y^{(m)} f_X^{(l)}$ $(l < m)$ indiquent qu'un sujet tiré au sort de l'échantillon **X** est dans un niveau bas qu'un sujet tiré au sort de l'échantillon **Y**.
- La somme des éléments de la diagonale (la trace de la matrice $\boldsymbol{\omega_{(k,k)}}$) divisée par $n_x n_y$ est équivalent à (eq. 12).

2. On estime π^0 par $\hat{\pi}^0 = Tr(\boldsymbol{\omega})/n_x n_y$ où *Tr* est la trace.
3. Construisons la matrice $\boldsymbol{W_{(k,k)}}$ en multipliant les éléments digonaux $\boldsymbol{\omega_{(k,k)}}$ par zero ($x_i = y_j$, eq. 1), les éléménts au-dessus de la diagonale par 1 ($x_i > y_j$, eq. 1) et les éléménts au-dessous de la diagonale par -1 ($x_i < y_j$, eq. 1):

$$\boldsymbol{W} = \begin{pmatrix} 0 & f_Y^{(1)} f_X^{(2)} & \cdots & f_Y^{(1)} f_X^{(k-1)} & f_Y^{(1)} f_X^{(k)} \\ -f_Y^{(2)} f_X^{(1)} & 0 & \cdots & f_Y^{(2)} f_X^{(k-1)} & f_Y^{(2)} f_X^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -f_Y^{(k-1)} f_X^{(1)} & -f_Y^{(k-1)} f_X^{(2)} & \cdots & 0 & f_Y^{(k-1)} f_X^{(k)} \\ -f_Y^{(k)} f_X^{(1)} & -f_Y^{(k)} f_X^{(2)} & \cdots & -f_Y^{(k)} f_X^{(k-1)} & 0 \end{pmatrix}$$

4. Calculons W (eq. 10) comme etant la somme de toous les éléments $\boldsymbol{W_{(k,k)}}$.
5. Estimons $\hat{\pi}^+$ comme etant la somme des éléments au-dessus de la diagonale de $\boldsymbol{W}$ divisée par $n_x n_y$.
6. Estimons $\hat{\pi}^-$ comme etant la valeur absolue de la somme des éléments au-dessous de la diagonale de $\boldsymbol{W}$ divisée par $n_x n_y$.

Notons que $\hat{\pi}^-, \hat{\pi}^+, \hat{\pi}^0$ et W estimés dans les étapes 2 and 4-6 vérifient les équations (eq.13) et (eq.14).

Une fois π^+ et π^- sont estimés, la méthode Improved RIDIT permet de calculer les statistiques suivantes :

2.1. The Absolute Risk Reduction (ARR) and Needed Number to Treat (NNT)

Considérons le cas de l'étude de l'impact d'un traitement où X et Y sont deux échantillons indépendants représentant deux groupes de sujets Control et Traitement. Si le traitement a un effet positif, alors π^- = P(Y<X) indique la probabilité qu'un sujet du groupe traitement est dans un état meilleur qu'un sujet du groupe control et π^+ = P(Y>X) indique qu'un sujet du groupe traitement est dans un état médiocre qu'un sujet du groupe control. Shepstone (Shepstone, 2001) a montré que P(Y<X) - P(Y>X) est équivalent à Absolute Risk Reduction (ARR). En utilisant de la méthode Improved RIDIT, on estime aisément ARR par :

$$\widehat{ARR} = \hat{\pi}^- - \hat{\pi}^+ \ (eq.16)$$

Aussi, on estime le Needed Number to Treat (NNT) qui est très important dans les essais cliniques pour évaluer l'efficacité d'un traitement. Le NNT est défini comme le nombre de patients à traiter pour prévenir un effet négatif (Wen et al. 2005). Il est estimé par NNT = 100/ARR (où ARR est en %) ou NNT = 1/ARR (où ARR est une proportion).

Dans le cas des études portant sur l'évaluation de l'effet négatif d'un évènement (exemple, maladie), on définit Absolute Risk Increase (ARI) par analogie à ARR dans les études cliniques. Le ARI est estimé par :

$$\widehat{ARI} = \hat{\pi}^+ - \hat{\pi}^- \ (eq.17)$$

L'intervalle de confiances de ARR/ARI à (1-α) est donné par :

$$CI_{1-\frac{\alpha}{2}} = \frac{|W|}{n_y n_x} \pm \frac{Z_{\frac{\alpha}{2}}}{n_y n_x} \sqrt{Var(W)} \ (eq.18)$$

2.2. The odds ordinal

Agresti (Agresti, 1980) a proposé la généralisation de l'odds ratio aux variables ordinales par :

$$odds_{ordinal} = \frac{P(Y > X)}{P(X > Y)} \quad (eq. 19)$$

La méthode Improved RIDIT permet de calculer cette statistique par :

$$\widehat{odds}_{ordinal} = \frac{\hat{\pi}^{+}}{\hat{\pi}^{-}} \quad (eq. 20)$$

3. Application aux données de la qualité de vie liée à la santé

Nous avons appliqué la méthode Improved RIDIT sur les données de la qualité de vie de 1857 sujets (1016 femmes and 841 hommes) diabétiques évalués par l'instrument EQ-5D-5L qui permet de mesurer la qualité de vie liée à la santé par 5 dimensions de la santé (Mobilité, Activités courantes, Autonomie, Douleur et Anxieté/Depression) par l'utilisation d'une échelle Likert à 5 niveaux (Mateo et al., 2015). En appliquant la méthode Improved RIDIT sur ces données, nous avons estimé le ARI et le $odds_{Ordinal}$ (**Tableau 2**). En terme de risque, les valeurs ARI ont montré que les femmes diabétiques avaient un risque absolu de diminuer leur mobilité de 22%, leur autonomie par 16% et leur activités courantes par 23%. Ces femmes diabétiques avaient aussi un niveau de douleur plus élevé de 29% par rapport aux hommes diabétiques et elles étaient 20% plus anxieuses. Les valeurs de $odds_{Ordinal}$ ont révélé les mêmes résultats en montrant que le diabète affecte la mobilité, l'autonomie, les activités courantes, l'anxiété et la douleur par deux fois chez les femmes que chez les hommes (**Tableau 2**).

Tableau 2. Comparaison entre la qualité de vie des femmes et des hommes diabétiques en appliquant la méthode Improved RIDIT aux données de Mateo et al.

EQ-5D dimension	**ARI [CI_{min} – CI_{max}]**	**$Odds_{Ordinal}$ [CI_{min} – CI_{max}]**	**W-test (P_{HB}-*value*)**
Mobilité	0.22 [0.17 - 0.27]	2.00 [1.76-2.28]	8.88 (< 0.0001)
Autonomie	0.16 [0.12 - 0.20]	2.33 [2.07-2.73]	8.12 (< 0.0001)
Activités courantes	0.23 [0.19 - 0.28]	2.35 [2.07-2.67]	9.90 (< 0.0001)
Douleur	0.29 [0.24 - 0.34]	2.38 [2.10-2.71]	11.35 (< 0.0001)
Anxiété/Depression	0.20 [0.15 - 0.24]	2.54 [2.15-2.80]	9.07 (< 0.0001)

P_{HB}-*value* the Bonferroni-Holm *p-values* correction.

4. Conclusion

La méthode Improved RIDIT permet de comparer entre deux échantillons indépendants en fournissant 3 informations statistiques : OddsOrdinal, Absolute Risk Increase/Absolute Risk Reduction et le Number Needed to Treat. Elle peut être utilisée dans plusieurs domaines de la recherche quand la variable à l'étude est de type ordinal. Dans le domaine de la santé, la méthode est utilisable dans les essais cliniques et dans l'étude de des maladies sur l'état de santé des populations.

Bibliographie

[1] Marfak A, Youlyouz-Marfak Y, El Achhab Y, Saad E, Nejjari C, Hilali A, Turman J. Improved RIDIT statistic approach provides more intuitive and informative interpretation of EQ-5D data, Health Qual Life Outcomes. 18 (2020). doi.org/10.1186/s12955-020-01313-3.
[2] Shepstone L. Re-conceptualising and generalising the Absolute Risk Difference: A unification of effect sizes, odds ratios and Number-Needed-to-Treat. Journal of Epidemiology & Community Health. 2001; 55:A7.
[3] Wen L, Badgett R, Cornell J. Number needed to treat: a descriptor for weighing therapeutic options. Am J Health Syst Pharm. 2005; 62(19):2031-2036.
[4] Agresti A. Generalized odds ratios for ordinal data. Biometrics. 1980; 36(1):59-67.
[5] Mateo DC, Gordillo MG, Olivares PR, Adsuar JC. Normative values of EQ-5D-5L for diabetes patients from Spain. Nutr Hosp. 2015; 32(4):1595-1602.

Sur une généralisation de la méthode PCO

Nicolas MARIE [1]

[1] *Laboratoire Modal'X, Université Paris Nanterre, Nanterre, France*
nmarie@parisnanterre.fr

Résumé. Pour le modèle de régression $Y = b(X)+\sigma(X)\varepsilon$, où la loi de X a une densité f, l'exposé portera sur une inégalité d'oracle pour un estimateur de bf, faisant intervenir un noyau au sens de Lerasle et al. (2016), sélectionné via la méthode PCO. En plus de la sélection de fenêtre pour des estimateurs à noyaux (au sens usuel) comme dans Lacour, Massart et Rivoirard (2017) ou Comte et Marie (2020), la méthode couvre la sélection de dimension pour des estimateurs par projection de f et bf dans le cas anisotrope (cf. Halconruy et Marie (2020)).

Mots-clés. Estimateurs non-paramétriques ; Estimateurs par projection ; Sélection de modèle ; Régression.

Abstract. In the regression model $Y = b(X) + \sigma(X)\varepsilon$, where X has a density f, this talk deals with an oracle inequality for an estimator of bf, involving a kernel in the sense of Lerasle et al. (2016), selected via the PCO method. In addition to the bandwidth selection for kernel-based estimators as in Lacour, Massart and Rivoirard (2017) or Comte and Marie (2020), the dimension selection for anisotropic projection estimators of f and bf is covered (see Halconruy and Marie (2020)).

Keywords. Nonparametric estimators ; Projection estimators ; Model selection ; Regression model.

Résumé détaillé

Soient $n \in \mathbb{N}^*$ variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$ ($d \in \mathbb{N}^*$), de même loi de probabilité absolument continue par rapport à la mesure de Lebesgue, et

$$\widehat{s}_{K,\ell}(n; x) := \frac{1}{n} \sum_{i=1}^{n} K(X_i, x)\ell(Y_i) \ ; \ x \in \mathbb{R}^d,$$

où $\ell : \mathbb{R} \to \mathbb{R}$ est une fonction borélienne et K est une application symétrique de $\mathbb{R}^d \times \mathbb{R}^d$ dans $\mathbb{R}$. Il s'agit d'un estimateur de la fonction $s : \mathbb{R}^d \to \mathbb{R}$ définie par

$$s(x) := \mathbb{E}(\ell(Y_1)|X_1 = x)f(x) \ ; \ \forall x \in \mathbb{R}^d,$$

où f est une densité de X_1. Pour $\ell = 1$, $\widehat{s}_{K,\ell}(n; .)$ est l'estimateur de f étudié dans Lerasle et al. [10], généralisant l'estimateur de Parzen-Rosenblatt et les estimateurs par

projection (cf. Parzen [12], Rosenblatt [13], Tsybakov [14], etc.), mais pour $\ell \neq 1$, il généralise des estimateurs utilisés en régression non-paramétrique. Supposons que pour tout $i \in \{1, \ldots, n\}$,

$$Y_i = b(X_i) + \sigma(X_i)\varepsilon_i \tag{1}$$

où ε_i est une variable aléatoire centrée et de variance 1, indépendante de X_i, et $b, \sigma : \mathbb{R}^d \to \mathbb{R}$ sont des fonctions boréliennes.

- Si $\ell = \mathrm{Id}_{\mathbb{R}}$, k est un noyau symétrique et

$$K(x', x) = \prod_{q=1}^{d} \frac{1}{h_q} k\left(\frac{x'_q - x_q}{h_q}\right) \text{ avec } h_1, \ldots, h_d > 0 \tag{2}$$

  pour tous $x, x' \in \mathbb{R}^d$, alors $\widehat{s}_{K,\ell}(n; .)$ est le numérateur de l'estimateur de Nadaraya-Watson de la fonction de régression b (cf. Nadaraya [11] et Watson [16]). Plus précisément, $\widehat{s}_{K,\ell}(n; .)$ est un estimateur de $s = bf$ car ε_1 est indépendante de X_1 et $\mathbb{E}(\varepsilon_1) = 0$. Si $\ell \neq \mathrm{Id}_{\mathbb{R}}$, alors $\widehat{s}_{K,\ell}(n; .)$ est le numérateur de l'estimateur étudié dans Einmahl et Mason [4, 5].

- Si $\ell = \mathrm{Id}_{\mathbb{R}}$, $\mathcal{B}_{m_q} = \{\varphi_1^{m_q}, \ldots, \varphi_{m_q}^{m_q}\}$ ($m_q \in \mathbb{N}^*$ et $q \in \{1, \ldots, d\}$) est une famille orthonormée de $\mathbb{L}^2(\mathbb{R})$ et

$$K(x', x) = \prod_{q=1}^{d} \sum_{j=1}^{m_q} \varphi_j^{m_q}(x_q)\varphi_j^{m_q}(x'_q) \tag{3}$$

  pour tous $x, x' \in \mathbb{R}^d$, alors $\widehat{s}_{K,\ell}(n; .)$ est un estimateur par projection sur $\mathcal{S} = \mathrm{span}(\mathcal{B}_{m_1} \otimes \cdots \otimes \mathcal{B}_{m_d})$ de $s = bf$.

Enfin, si $b = 0$ dans le Modèle (1), pour tout $i \in \{1, \ldots, n\}$,

$$Y_i = \sigma(X_i)\varepsilon_i \tag{4}$$

Si $\ell(x) = x^2$ pour tout $x \in \mathbb{R}$, alors $\widehat{s}_{K,\ell}(n; .)$ est un estimateur de $s = \sigma^2 f$.

Ces dernières années, plusieurs méthodes de sélection de fenêtre pour l'estimateur de Parzen-Rosenblatt ($\ell = 1$ et K défini par (2)) ont été étudiées. D'une part, la méthode de Goldenshluger-Lepski, introduite dans [6], très satisfaisante sur le plan théorique, mais pas totalement sur le plan numérique (cf. Comte and Rebafka [3]). D'autre part, dans [9], Lacour, Massart et Rivoirard ont proposé la méthode PCO (Penalized Comparison to Overfitting) et démontré une inégalité d'oracle en usant d'une inégalité de concentration pour les U-statistiques due à Houdré et Reynaud-Bouret [8]. Avec Varet, les auteurs de [9] ont établi l'efficacité de la méthode PCO sur le plan numérique dans Varet et al. [15]. Toujours dans le contexte de l'estimation de densité, la méthode PCO a été étendue à

la sélection de fenêtres pour l'estimateur récursif de Wolverton-Wagner dans Comte et Marie [1].

L'exposé portera sur l'extension suivante de la méthode PCO à la sélection de l'application symétrique K pour l'estimateur $\widehat{s}_{K,\ell}(n;.)$:

$$\widehat{K} \in \arg\min_{K\in\mathcal{K}_n}\left\{\|\widehat{s}_{K,\ell}(n;.)-\widehat{s}_{K_0,\ell}(n;.)\|_2^2+\frac{2}{n^2}\sum_{i=1}^{n}\langle K(.,X_i),K_0(.,X_i)\rangle_2\ell(Y_i)^2\right\}$$

où

$$K_0 \in \arg\max_{K\in\mathcal{K}_n}\left\{\sup_{x\in\mathbb{R}^d}|K(x,x)|\right\}$$

est un *noyau maximal* et $\mathcal{K}_n$ désigne une famille d'applications symétriques, par exemple de la forme (2) ou de la forme (3). Quelques expériences numériques seront présentées, puis il sera établi que si $\mathbb{E}(\exp(\alpha|\ell(Y_1)|))<\infty$, alors

$$\begin{aligned}\mathbb{E}(\|\widehat{s}_{\widehat{K},\ell}(n;.)-s\|_2^2) \;\leqslant\; (1+\theta)\min_{K\in\mathcal{K}_n}\mathbb{E}(\|\widehat{s}_{K,\ell}(n;.)-s\|_2^2)\\ +\frac{\mathfrak{c}}{\theta}\left(\|\mathbb{E}(\widehat{s}_{K_0,\ell}(n;.))-s\|_2^2+\frac{\log(n)^5}{n}\right).\end{aligned}$$

Pour le modèle (1), l'exposé s'achèvera sur une borne de risque pour l'estimateur quotient

$$\frac{\widehat{s}_{\widehat{K}_1,\mathrm{Id}_\mathbb{R}}(n;.)}{\widehat{s}_{\widehat{K}_2,1}(n;.)}\text{ de }b(.),$$

où

$$\widehat{K}_1 \in \arg\min_{K\in\mathcal{K}_n}\left\{\|\widehat{s}_{K,\mathrm{Id}_\mathbb{R}}(n;.)-\widehat{s}_{K_0,\mathrm{Id}_\mathbb{R}}(n;.)\|_2^2+\frac{2}{n^2}\sum_{i=1}^{n}\langle K(.,X_i),K_0(.,X_i)\rangle_2 Y_i^2\right\}$$

et

$$\widehat{K}_2 \in \arg\min_{K\in\mathcal{K}_n}\left\{\|\widehat{s}_{K,1}(n;.)-\widehat{s}_{K_0,1}(n;.)\|_2^2+\frac{2}{n^2}\sum_{i=1}^{n}\langle K(.,X_i),K_0(.,X_i)\rangle_2\right\}.$$

Ces résultats sont issus de deux travaux, l'un en collaboration avec Fabienne Comte [2], et l'autre avec Hélène Halconruy [7].

References

[1] F. Comte et N. Marie. *Bandwidth Selection for the Wolverton-Wagner Estimator.* Journal of Statistical Planning and Inference 207, 198-214, 2020.

[2] F. Comte et N. Marie. *On a Nadaraya-Watson Estimator with Two Bandwidths.* A paraître dans Electronic Journal of Statistics, 2021.

[3] F. Comte et T. Rebafka. *Nonparametric Weighted Estimators for Biased Data.* Journal of Statistical Planning and Inference 174, 104-128, 2016.

[4] U. Einmahl et D.M. Mason. *An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators.* Journal of Theoretical Probability 13, 1-37, 2000.

[5] U. Einmahl et D.M. Mason. *Uniform in Bandwidth Consistency of Kernel-Type Function Estimators.* Annals of Statistics 33, 1380-1403, 2005.

[6] A. Goldenshluger et O. Lepski. *Bandwidth Selection in Kernel Density Estimation: Oracle Inequalities and Adaptive Minimax Optimality.* The Annals of Statistics 39, 1608-1632, 2011.

[7] H. Halconruy et N. Marie. *Kernel Selection in Nonparametric Regression.* A paraître dans Mathematical Methods of Statistics, 2021.

[8] C. Houdré et P. Reynaud-Bouret. *Exponential Inequalities, with Constants, for U-statistics of Order Two.* Stochastic Inequalities and Applications, vol. 56 of Progr. Proba., 55-69, Birkhauser, 2003.

[9] C. Lacour, P. Massart et V. Rivoirard. *Estimator Selection: a New Method with Applications to Kernel Density Estimation.* Sankhya A 79, 2, 298-335, 2017.

[10] M. Lerasle, N.M. Magalhaes et P. Reynaud-Bouret. *Optimal Kernel Selection for Density Estimation.* High dimensional probabilities VII: The Cargese Volume, vol. 71 of Prog. Proba., 435–460, Birkhauser, 2016.

[11] E.A. Nadaraya. *On a Regression Estimate.* (Russian) Verojatnost. i Primenen. 9, 157-159, 1964.

[12] E. Parzen. *On the Estimation of a Probability Density Function and the Mode.* The Annals of Mathematical Statistics 33, 1065-1076, 1962.

[13] M. Rosenblatt. *Remarks on some Nonparametric Estimates of a Density Function.* The Annals of Mathematical Statistics 27, 832-837, 1956.

[14] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

[15] S. Varet, C. Lacour, P. Massart et V. Rivoirard. *Numerical Performance of Penalized Comparison to Overfitting for Multivariate Density Estimation.* Preprint, 2020.

[16] G.S. Watson. *Smooth Regression Analysis.* Sankhya A 26, 359-372, 1964.

Quelques tests de détection exploitant des données d'apprentissage en astronomie

David Mary[1], Étienne Roquain[2], Sophia Sulis[3] & Sébastien Bourguignon[4]

[1] *Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Boulevard de l'Observatoire, CS 34229, 06304, Nice cedex 4, France; David.Mary@unice.fr*

[2] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001), Faculté des Sciences et Ingénierie, Tour 15/16, Étage 2, BC 158, 4, place Jussieu, 75252, PARIS cedex 05; Etienne.Roquain@upmc.fr*

[3] *Aix Marseille Université, CNRS, CNES, LAM, Marseille, France; Sophia.Sulis@lam.fr*

[4] *Laboratoire des Sciences du Numérique de Nantes (LS2N, UMR 6004), École Centrale de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3; Sebastien.Bourguignon@ec-nantes.fr*

Résumé. Cette note est une version actualisée de la contribution proposée aux JdS 2020 qui ont été reportées en 2021. Nous considérons le problème de construction de tests de détection pour certains types de données issues de l'astrophysique. Une caractéristique commune des données considérées est que la distribution du bruit de fond est inconnue, ce qui invalide l'utilisation de nombreux tests classiques. Nous présentons des résultats récents proposant des solutions à ce problème pour deux applications spécifiques, la détection d'exoplanètes et celle de galaxies. Dans les deux cas, nous exploitons pour contôler les taux d'erreur le fait que des données d'apprentissage sont disponibles.

Mots-clés. détection, tests multiples, exoplanètes, galaxies.

Abstract.

This note is an updated version of the contribution proposed to the JdS 2020 conference that was postponed to 2021. In this note, we consider the problem of building detection tests for some types of astrophysical data. The distribution of the background noise is unknown in the considered cases, preventing from using most classical procedures. We present works that address this problem for two specific applications : the detection of exoplanets and of galaxies. In both case, we exploit the availability of training data sets to allow for the control of the test's error rates.

Keywords. detection, multiple testing, exoplanets, galaxies.

1 Problématique et contexte

La détection de sources est un enjeu majeur dans plusieurs domaines de l'astrophysique. Le contexte moderne d'instruments toujours plus complexes produisant des données toujours plus volumineuses conduit souvent à réaliser un nombre gigantesque de tests simultanément. La thématique des tests multiples permet de prendre en compte cette multiplicité de manière appropriée. Si ce domaine de recherche possède des origines très anciennes en statistique, les deux dernières décennies y ont vu une explosion de travaux théoriques et appliqués : parmi les procédures qui ont vu le jour, on peut citer notamment la procédure de Benjamini-Hochberg [1] qui contrôle le *false discovery rate* (FDR), ainsi que les différentes versions du *higher criticism* (HC) [2] et des tests de Berk-Jones (BJ) [3] pour le cas de signaux rares et faibles [2].

En astrophysique, ce genre de procédures reste cependant assez peu utilisé. Une des raisons est que la distribution des données sous l'hypthèse nulle est très souvent inconnue. Ceci pousse l'utilisateur à se tourner vers des procédures *ad hoc*, qui peuvent présenter des risques. Un exemple récent est le cas de la détection d'une exoplanète près de l'étoile α Centauri Bb [4]. La détection de cette exoplanète, annoncée en 2012 sur la base d'une P-valeur évaluée à 0.02% [4], a été fortement remise en cause depuis [5, 6]. En effet, ces dernières analyses ont mis en lumière des effets mal pris en compte, en particulier le signal parasite émis par l'étoile elle-même (le "bruit stellaire"). Ces effets, très difficiles à contrôler, impactent fortement les taux d'erreur prédits par les tests qui les ignorent.

L'objet de cette note est de présenter des procédures de détection avec des taux d'erreur correctement contrôlés même lorsque la distribution des statistiques de test sous la distribution nulle est mal connue. Nous considérons deux cas concrets : la détection d'exoplanètes par vélocimétrie radiale [7] (Section 2), et la détection de galaxies dans des images multi-longueurs d'onde de l'instrument MUSE [8, 9] (Section 3).

2 Détection d'exoplanètes par vitesses radiales

La présence d'une planète orbitant autour d'une étoile induit un mouvement de l'étoile autour du barycentre de masse du système étoile-planète. Ce mouvement module de façon quasi-périodique la vitesse radiale de l'étoile par rapport à un observateur terrestre et s'imprime par effet Doppler sur la lumière de l'étoile. En mesurant le décalage Doppler des raies du spectre stellaire en fonction du temps, on déduit ainsi la vitesse radiale de l'étoile. Les données se présentent donc sous la forme d'une série temporelle obtenue durant une fenêtre temporelle d'observation.

Le but est de tester, dans une telle série de vitesses radiales, H_0 : la moyenne est nulle (il n'y a pas d'exoplanète) contre H_1 : la moyenne est un signal quasi-périodique (il y a une ou plusieurs exoplanètes). La série temporelle est supposée stationnaire, mais sa matrice de covariance Σ est inconnue en raison du bruit stellaire. Pour s'adapter à l'alternative, une approche classique en échantillonnage régulier consiste à construire le

périodogramme des données (module carré de la transformée de Fourier discrète de la série), puis à rechercher si ses composantes indiquent de façon significative la présence de signaux périodiques. Cependant, la distribution de ce périodogramme est inconnue sous H_0 car Σ est inconnue.

Pour contourner cette difficulté nous avons proposé [10] une approche "exogène", rendue possible par la disponibilité d'un certain nombre (L) de séries simulées sous H_0. En effet, il existe à l'heure actuelle des codes astrophysiques permettant de simuler de façon réaliste les vitesses radiales qui seraient observées pour un type d'étoile donné. Ceci permet, au moins sur une certaine plage de fréquences, de générer des séries exogènes sous H_0. Celles-ci sont cependant en nombre très limité (quelques unités à une dizaine) en raison du fort coût de calcul que ces simulations nécessitent. A l'aide de ces séries, il est possible de construire un périodogramme de référence, utilisé pour normaliser le périodogramme des données.

Notons $Z_1, \ldots, Z_N$ un sous-ensemble de composantes du périodogramme ainsi standardisé. Si l'échantillonnage est régulier, alors les distributions considérées possèdent des expressions analytiques [11]. Ainsi, sous certaines conditions, on peut montrer que $Z_1, \ldots, Z_N$ sont asymptotiquement indépendants avec $Z_j \sim \mathcal{F}(2, 2L)$ (loi de Fisher de paramètres 2 et $2L$) sous l'hypothèse nulle et $Z_j \sim \mathcal{F}_{\lambda_j}(2, 2L)$ (loi de Fisher decentrée avec un certain paramètre de décentrage λ_j et de paramètres 2 et $2L$) sous l'alternative. Soulignons que sous H_0, la distribution du périodogrammee standardisé est indépendante du paramètre de nuisance Σ. Cette propriété, complètée par celle de l'indépendance asymptotique, permet de construire des tests globaux de niveau α à partir des Z_j. Par exemple, une façon naturelle et explicite d'agréger ces tests est simplement le test du maximum. Dans le cas considéré, celui-ci rejette l'hypothèse nulle si $\max(Z_j, 1 \leq j \leq N)$ dépasse le seuil $\gamma(\alpha)$, calibré de sorte que

$$1 - \Big(1 - \Big(\frac{L}{\gamma(\alpha) + L}\Big)^L\Big)^N = \alpha.$$

Des résultats similaires peuvent être obtenus pour de nombreux autres tests comme celui de Fisher [12], ses variantes [13, 14, 15] et ceci permet également d'utiliser des tests de détection de type HC et BJ.

Sur des simulations, on voit que l'approximation asymptotique est bonne quand la longueur de la série est suffisamment grande devant la durée caractéristique de corrélation du signal (typiquement un à deux ordres de grandeur supérieur), ce qui confirme que le test conduit est bien de niveau α. De plus, les modèles analytiques obtenus permettent d'étudier la puissance asymptotique de cette procédure de test. Là aussi, les simulations montrent que ces expressions sont valables pour des valeurs de N modérées . Ces résultats permettent de faire des études de détectabilité pour des planètes ou des instruments avec des caractéristiques données.

Dans le cas de l'échantillonnage irrégulier, l'hypothèse d'indépendance entre les Z_j ne peut être tenue. Nous avons proposé dans [16] des techniques de *bootstrap* pour approcher

la distribution de la statistique de test sous H_0. Ces techniques permettent d'estimer, par des méthodes de Monte Carlo, la probabilité de fausse alarme $\alpha(\gamma)$ (aussi notée PFA dans la Figure 1) pour tout seuil γ et pour toute grille d'échantillonnage donnés. Deux exemples sont montrés en Figure 1, où 100 estimés de PFA ont ainsi été générés par *bootstrap* pour deux types d'échantillonnages (irrégulier et régulier avec données manquantes sur une certaine durée, en gris). On voit que les vraies valeurs $\alpha(\gamma)$ de la procédure de détection proposée (en vert) sont comprises dans l'intervalle obtenu par la procédure d'estimation de $\alpha(\gamma)$ pour chaque γ.

3 Détection de galaxies dans les données MUSE

Le spectrographe intégral de champ MUSE installé sur un des télescopes de 8 m au Very Large Telescope (Chili) permet d'obtenir des images multi-longueurs d'onde (typiquement 300×300 images dans 3600 canaux optiques). On cherche à détecter dans ce "cube" de données des galaxies très lointaines, qui se manifestent par une faible augmentation locale du flux (une raie en émission) dans une poignée de voxels. On connaît à peu près la forme de ces raies mais ni leurs hauteurs, ni leur nombre (quelques dizaines à quelques centaines typiquement) ni leurs positions dans le cube. Par ailleurs, la hauteur de certaines raies peut être beaucoup plus faible que le niveau du bruit de fond et que celui d'autres sources brillantes et étendues (d'autres étoiles et galaxies) voire que certains artefacts instrumentaux. On regroupe ces sources parasites sous le terme de signaux de nuisance.

Dans ce cadre, nous avons proposé dans [9] une approche de détection en deux temps : les signaux de nuisance sont d'abord supprimés et l'étape de détection des galaxies se fait ensuite, dans le résidu. Pour s'adapter à l'alternative, l'approche considère les maximas locaux du cube de données résiduel. En raison des prétraitements, la distribution sous H_0 des maxima locaux n'est pas connue. Le problème considéré ici est plus complexe que le précédent puisqu'il y a plusieurs hypothèses nulles, chacune liée à un maximum local [17]. Si nous notons x, y, z la position d'un maximum local, on teste $H_{0,x,y,z}$: il n'y a pas de raie d'émission à la position (x, y, z), contre $H_{1,x,y,z}$: il y en a une à cette position. Le critère d'erreur considéré est celui du FDR, moyenne du taux de fausses découvertes parmi les positions déclarées comme correspondant à une galaxie.

L'approche que nous proposons pour ce problème est "endogène". Dans [9] nos simulations numériques suggèrent que la procédure proposée, qui utilise la population des minimas locaux comme "proxy" pour celle des maxima locaux, permet de contrôler le FDR. Dans cette contribution, nous démontrons mathématiquement qu'une procédure voisine contrôle le FDR sous plusieurs hypothèses simplificatrices.

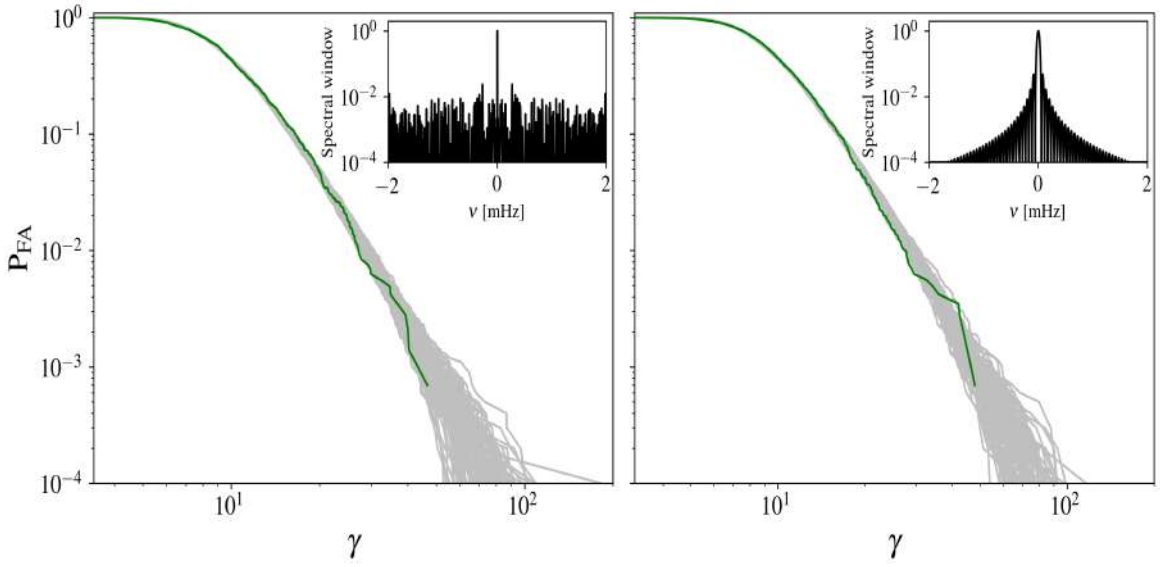


Figure 1: Probabilité de fausse alarme α (notée PFA) en fonction du seuil de détection (γ) pour deux types d'échantillonnages irréguliers : un échantillonnage aléatoire (10% des observations sont prises aléatoirement dans une grille régulière de 2880 points, panel de gauche) et un échantillonnage régulier "avec trou" (pas régulier de $\Delta t = 60$ s mais absence de données pendant 43.2 h résultant également en $N = 288$ points, panel de droite). Les fenêtres spectrales associées à chaque type d'échantillonnage sont montrées en encart dans chaque panel. Les données utilisées pour cette validation sont des séries temporelles de vitesses radiales du Soleil obtenues par l'instrument GOLF sur le satellite SoHo, et les signaux d'apprentissage utilisés pour la calibration du périodogramme proviennent de simulations magnéto-hydrodynamiques (MHD) de convection solaire [10]. Les courbes vertes montrent la valeur $\alpha(\gamma)$, mesurée sur les données GOLF, de la procédure de test exploitant les signaux d'apprentissage. Les courbes grises montrent les estimés $\widehat{\alpha}(\gamma)$ tels qu'obtenus par notre technique de *bootstrap*. On voit que cette technique permet d'estimer de façon fiable le taux d'erreur $\alpha(\gamma)$ de la procédure de test proposée.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B, 57(1):289–300, 1995.

[2] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat., 2004.

[3] V. Gontscharuk et al. The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. Biom. J., 57(1):159–180, 2014.

[4] X. Dumusque et al. An Earth-mass planet orbiting α Centauri B. Nature, 491:207–211, 2012.

[5] A. Hatzes. The Radial Velocity Detection of Earth-mass Planets in the Presence of Activity Noise: The Case of α Centauri Bb. ApJ, 770:133, 2013.

[6] V. Rajpaul et al. Ghost in the time series: no planet for Alpha Cen B. MNRAS, 456:L6–L10, 2016.

[7] S. Sulis. Statistical methods using hydrodynamic simulations of stellar atmospheres for detecting exoplanets in radial velocity data. Thèse, Université Côte d'Azur, October 2017.

[8] R. et al. Bacon. The muse hubble ultra deep field survey - i. survey description, data reduction, and source detection. A&A, 608:A1, 2017.

[9] D. Mary, R. Bacon, S. Conseil, L. Piqueras, and A. Schutz. Origin: Blind detection of faint emission line galaxies in muse datacubes. Astronomy Astrophysics, Jan 2020.

[10] S. Sulis, D. Mary, and L. Bigot. 3D magneto-hydrodynamical simulations of stellar convective noise for improved exoplanet detection. I. Case of regularly sampled radial velocity observations. A&A, 635:A146, March 2020.

[11] S. Sulis, D. Mary, and L. Bigot. A study of periodograms standardized using training datasets and application to exoplanet detection. IEEE Transactions on Signal Processing, 65(8):2136–2150, April 2017.

[12] R.A. Fisher. Tests of Significance in Harmonic Analysis. Proc. R. Soc. London, Ser. A, 125:54–59, 1929.

[13] S.T. Chiu. Detecting periodic components in a white gaussian time series. J. R. Stat. Soc. Series B, 51(2):249–259, 1989.

[14] M. Shimshoni. On fisher's test of significance in harmonic analysis. Geophys. J. R. Astronom. Soc., pages 373–377, 1971.

[15] E. Bölviken. New tests of significance in periodogram analysis. Scandinavian J. Stat., 10(1):1–9, 1983.

[16] S. Sulis, D. Mary, and Lionel Bigot. A bootstrap method for sinusoid detection in colored noise and uneven sampling. application to exoplanet detection. In EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017, pages 1095–1099. IEEE, 2017.

[17] Dan Cheng and Armin Schwartzman. Multiple testing of local maxima for detection of peaks in random fields. Ann. Statist., 45(2):529–556, 04 2017.

Une approche de régularisation itérative pour fonctions convexes

Mathurin Massias [1] & Cesare Molinari [2] & Lorenzo Rosasco [3] & Silvia Villa [4]

[1] *Malga, DIBRIS, University of Genova*
[2] *Istituto Italiano di Tecnologia, Genova*
[3] *Malga, DIBRIS, University of Genova & MIT*
[4] *Malga, DIMA, University of Genova*

Résumé. Nous proposons une approche de régularisation itérative pour les modèles linéaires, quand le biais imposé à la solution n'est pas fortement convexe. Nous caractérisons la stabilité d'une approche primale-duale en présence de bruit déterministique. Les résultats théoriques sont complétés par des expériences montrant des gains de temps de plusieurs ordres de grandeur.

Mots-clés. Problèmes inverses, régularisation, parcimonie.

Abstract. We study iterative/implicit regularization for linear models, when the bias is convex but not necessarily strongly convex. We characterize the stability properties of a primal-dual gradient based approach, analyzing its convergence in the presence deterministic noise. Theoretical results are complemented by experiments showing that state-of-the-art performances can be achieved with considerable computational speed-ups.

Keywords. Inverse problems, regularisation, sparsity.

1 Introduction

When estimating parameters of a machine learning model, a classical way to achieve uniqueness and stability is to consider explicitly penalized objectives, leading to regularized empirical risk minimization. A more recent approach is based on directly exploiting an iterative optimization procedure for an unpenalized problem. This approach is known as implicit regularization [Gunasekar et al., 2017], early stopping [Yao et al., 2007, Raskutti et al., 2014] or iterative regularization [Engl et al., 1996, Kaltenbacher et al., 2008]. In this work, we are interested in iterative regularization procedures where the considered bias is not the Euclidean norm but rather a general convex functional. Such a procedure has been studied when the bias is strongly convex: linearized Bregman iterations (a.k.a. mirror descent) can be used [Burger et al., 2007, Gunasekar et al., 2018]. The general convex case, even for linear models, is much less understood.

We propose and study an efficient algorithm for general convex bias. We adapt the Chambolle and Pock (CP) algorithm with errors [Chambolle and Pock, 2011, Rasch and

Chambolle, 2020], and study its iterative regularization properties. In the setting of linear models with worst case errors, our analysis provides dimension free convergence and stability results in terms of Bregman divergence and approximate feasibility.

Notation. The set of integers from 1 to n is $[n]$. Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $J\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous. The subdifferential of J at $w \in \mathbb{R}^p$ is $\partial J(w)$. The Bregman divergence associated to J is denoted $D_J^\theta(w, w') \triangleq J(w) - J(w') - \langle \theta, w - w' \rangle$, where $\theta \in \partial J(w')$. The Fenchel-Legendre conjugate of f is $f^\star(\theta) \triangleq \sup_w \langle w, \theta \rangle - f(w)$. The indicator function $\iota_{\{\mathbf{y}\}}$ is equal to zero if the argument equals $\mathbf{y}$ and $+\infty$ otherwise. For a proper convex lower semicontinuous function J, $\mathrm{prox}_J(x) = \mathrm{argmin}_y\, J(y) + \|x - y\|^2/2$.

2 Over-parametrization and regularization

Let J be a regularization functional, used to impose a structure on the solution. We consider linear models of the form $\mathbf{y} = \mathbf{X}w$, when there are possible infinitely many solutions for a given $\mathbf{y}$. In that case, the typical Tikhonov regularization is

$$\min_w \frac{1}{2}\|\mathbf{y} - \mathbf{X}w\|^2 + \lambda J(w) \ . \tag{1}$$

We study the alternative of iterative regularization, where the problem solved is

$$\min_{w \in \mathbb{R}^p} J(w) \quad \text{s.t.} \quad \mathbf{X}w = \mathbf{y} \ , \tag{2}$$

and where the number of iterations k plays the role of a regularization parameter just like λ in Tikhonov regularization (or rather $1/\lambda$). Iterative regularization is particularly appealing in the large scale setting, where substantial computational savings are expected: early stopping needs a finite number of iterations, while Tikhonov regularization requires solving exactly Problem (1) for multiple values of λ. The results in Benning et al. [2016] and Gunasekar et al. [2018] exhibit an iterative regularization proceduring by showing that mirror descent solves Problem (2) under the key assumption that J is strongly convex. In this paper, we take steps to fill in this gap studying an efficient approach for which we characterize the iterative regularization properties.

3 Problem setting and proposed algorithm

In the following, $\mathbf{X}$ is an n by p matrix and $\mathbf{y}$ an n-dimensional vector. We consider the case where $\mathbf{y}$ is unknown, and a vector $\mathbf{y}^\delta$ is available such that $\|\mathbf{y} - \mathbf{y}^\delta\| \leq \delta$, where $\delta \geq 0$ can be interpreted as the noise level.

Assumption 1 *We assume that the bias $J : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, and lower semicontinuous. We also assume that Problem (2) has at least one solution (in particular the linear equation has at least one solution for the exact data $\mathbf{y}$).*

Note that we use a vectorial notation for simplicity but our results are dimension free and sharp for an infinite dimensional setting where $\mathbf{X}$ is a linear bounded operator between separable Hilbert spaces. Some examples of this setting are given in the sequel.

Example 2 (Sparse recovery) *Choosing $J = \|\cdot\|_1$ corresponds to finding the minimal ℓ_1-norm solution to a linear system, and in this case Problem (2) is known as Basis Pursuit [Chen et al., 1998], and Problem (1) as the Lasso [Tibshirani, 1996].*

Example 3 (Low rank matrix completion) *In several applications, such as recommendation systems, it is useful to recover a low rank matrix, starting from the observation of a subset of its entries [Candès and Recht, 2009]. A convex formulation is:*

$$\min_{W \in \mathbb{R}^{p_1 \times p_2}} \|W\|_* \quad s.t.\ W_{ij} = Y_{ij} \quad \forall (i,j) \in \mathcal{D} \ ,$$

where $\|\cdot\|_*$ is the nuclear norm and $\mathcal{D} \subset [p_1] \times [p_2]$ is the set of observed entries of the matrix Y. In that case, $\mathbf{X}$ is not a design matrix: it is a (self-adjoint and linear) masking operator from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^{p_1 \times p_2}$, such that $(XW)_{ij}$ has value W_{ij} if $(i,j) \in \mathcal{D}$ and 0 otherwise; the constraints write $\mathbf{X}W = \mathbf{X}Y$.

Example 4 (Total variation) *The problem of Total Variation [Rudin et al., 1992] is $\min_{W \in \mathbb{R}^{p_1 \times p_2}} \|\nabla W\|_1 s.t.\ \mathbf{X}W = Y$, where $\mathbf{X}$ is usually a blurring operator. The problem can be reformulated as: $\min_{\tilde{W}} \Omega(\tilde{W})$ s.t. $\tilde{\mathbf{X}}\tilde{W} = \tilde{Y}$, with $\tilde{W} = \begin{pmatrix} W \\ U \end{pmatrix}$, $\Omega(\tilde{W}) = \|U\|_1$, $\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & 0 \\ \nabla & -\mathrm{Id} \end{pmatrix}$ and $\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$.*

Now consider the following iterations, with initialization $w_0 \in \mathbb{R}^p, \theta_0 = \theta_{-1} \in \mathbb{R}^n$, and parameters τ, σ such that $\sigma\tau \|\mathbf{X}\|_{\mathrm{op}}^2 < 1$:

$$\begin{cases} w_{k+1} = \mathrm{prox}_{\tau J}(w_k - \tau \mathbf{X}^\top (2\theta_k - \theta_{k-1})) \ , \\ \theta_{k+1} = \theta_k + \sigma(\mathbf{X} w_{k+1} - \mathbf{y}) \ . \end{cases} \tag{3}$$

Proposition 5 *Under Assumption 1, the iterations (3) converge to a point $(w^\star, \theta^\star)$ such that $\mathbf{X}w^\star = \mathbf{y}$. Additionally, $w^\star$ is a minimizer of J amongst all interpolating solutions, meaning that it solves Problem (2).*

We prove Proposition 5 by casting (3) as an instance of the Chambolle-Pock algorithm [Chambolle and Pock, 2011] which solves $\min_w f(\mathbf{X}w) + g(w)$. Hence, for $f = \iota_{\{\mathbf{y}\}}$ and $g = J$, it can minimize a convex function on a set defined by linear equalities, as in Problem (2).

4 Theoretical analysis

As usual for this class of methods, called primal-dual, the Lagrangian is a useful tool to establish convergence results. The Lagrangian of Problem (2) is $\mathcal{L}(w, \theta) = J(w) + \langle \theta, \mathbf{X}w - y \rangle$, where $\theta \in \mathbb{R}^n$ is the dual variable. Under a technical condition, $w^\star$ is a solution of Problem (2) if and only if there exists a dual variable $\theta^\star$ such that $(w^\star, \theta^\star)$ is a saddle-point for the Lagrangian, namely, iff for every $(w, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$,

$$\mathcal{L}(w^\star, \theta) \leq \mathcal{L}(w^\star, \theta^\star) \leq \mathcal{L}(w, \theta^\star) \ . \tag{4}$$

First, we need to choose a suitable criterion to estimate the approximation properties of the iterates. In general, it is not reasonable to expect a rate of convergence for the distance between the iterates and the solution. Indeed, since the problem is only convex, it is well known that the convergence in distance can be arbitrarily slow. In Molinari et al. [2021], we explain why a reasonable choice is given by the duality gap together with the residual norm (respectively, $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ and $\|\mathbf{X}w_k - \mathbf{y}\|$). For these two quantities, we derive early-stopping bounds in the inexact case, i.e. when the accessible data is only $\mathbf{y}^\delta$ with $\|\mathbf{y}^\delta - \mathbf{y}\| \leq \delta$ (Proposition 6 and Corollary 7).

We now consider the iterates (w_k, θ_k), and their averaged versions $(\overline{w}_k, \overline{\theta}_k)$, obtained by applying iterations (3) to the noisy problem, where $\mathbf{y}$ is replaced by $\mathbf{y}^\delta$ with $\|\mathbf{y}^\delta - \mathbf{y}\| \leq \delta$. In Proposition 6, we derive early-stopping bounds for the iterates, in terms of duality gap $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ and residual norm $\|\mathbf{X}w_k - \mathbf{y}\|$. We highlight that, despite the error in the data $\mathbf{y}^\delta$, both quantities are defined in terms of $\mathbf{y}$ and hence related to the noiseless problem. In particular, $(w^\star, \theta^\star)$ is a saddle-point for the noiseless Lagrangian. We have the following estimates (see Molinari et al. [2021] for proofs).

Proposition 6 (Stability) *Under Assumption 1, let $\varepsilon \in (0, 1)$ and assume that the step-sizes are such that $\sigma\tau \leq \varepsilon / \|\mathbf{X}\|_{\text{op}}^2$. Then, for $z = (w, \theta)$ and $V(z) = \|w\|^2/2\tau + \|\theta\|/^2/2\sigma$,*

$$\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k) \leq \frac{1}{k} \left(\sqrt{V(z_0 - z^\star)} + \sqrt{2\sigma}\delta k \right)^2 \tag{5}$$

and

$$\|\mathbf{X}\overline{w}^k - \mathbf{y}\|^2 \leq \frac{2(1+\varepsilon)}{\sigma\varepsilon(1-\varepsilon)} \left[\sqrt{2\sigma V(z_0 - z^\star)}\delta + \frac{\sigma\varepsilon}{1-\varepsilon}\delta^2 + 2\sigma\delta^2 k + \frac{1}{k} V(z_0 - z^\star) \right]. \tag{6}$$

Corollary 7 (Early-stopping) *Under the assumptions of Proposition 6, choose $k = c/\delta$ for some $c > 0$. Then there exist constants C, C' and C'' such that*

$$\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k) \ \leq \ C\delta \quad \text{and} \quad \|\mathbf{X}\overline{w}^k - \mathbf{y}\|^2 \leq C'\delta + C''\delta^2 \ .$$

5 Empirical analysis on sparse recovery

Additional experiments are in Molinari et al. [2021]. In real settings, $w^\star$ and δ are unknown, and so is the stopping time. It can still be evaluated by cross validation or similar procedure, as is usually done for the optimal λ in explicit regularization.

The most popular regularization approach is to solve Problem (1) (here, the Lasso) for typically 100 values of λ geometrically chosen as $\lambda_t = 10^{-3t/99}\|\mathbf{X}^\top \mathbf{y}\|_\infty$ for $t = 0, \dots, 99$. In Figure 1 we compare the Lasso regularization path to the Basis Pursuit optimization path of the Chambolle-Pock algorithm. The dataset for this experiment is rcv1-train, with $(n, p) = (20\,242, 26\,683)$. The figure of merit is the prediction mean squared error on left out data, using 4-fold cross validation (dashed color lines), with the average across the folds in black. The horizontal line marks the λ (resp. the iteration k) for which the Lasso path (resp. the optimization path of iterations (3)) reaches its minimum MSE on the test fold.

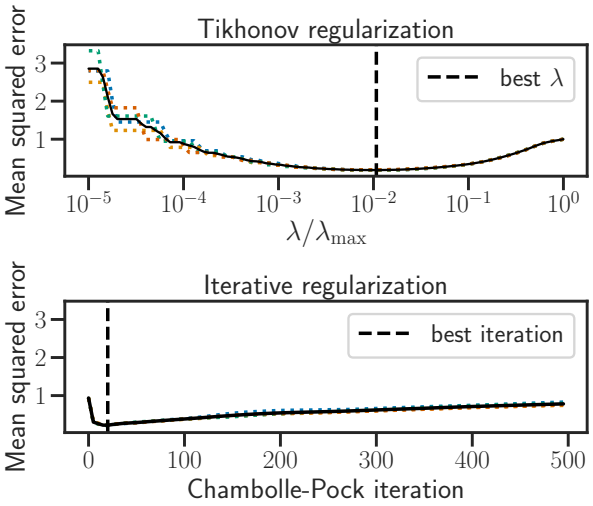


Figure 1: Comparison of Tikhonov regularization path (top) and optimization path of Algorithm (3) (bottom) on rcv1 with 4-fold cross-validation. Minimal value reached: 0.19 (top), 0.21 (bottom). Computation time up to optimal parameter: 50 s (top); 0.5 s (bottom).

The first observation is that the Basis Pursuit solution (both the end of the optimization ($k = +\infty$) and regularization paths ($\lambda = 0$)) performs very poorly, having a MSE greater than the one obtained by the 0 solution. The second observation is that the minimal MSEs on both paths are similar: 0.19 for Lasso path, 0.21 for optimization path of Algorithm (3). The main point is however that it takes 20 iterations of algorithm (3) to reach its best iterate, while the optimal λ for the Lasso is around $\lambda_{\max}/100$. If the default grid of 100 values between $\lambda_{\max}$ and $\lambda_{\max}/1000$ was used, this means that 66 Lassos must be solved, each one needing hundreds or thousands of iterations to converge. This is reflected in the timings: 0.5 s for Algorithm (3) vs 50 s for Tikhonov.

References

Martin Benning, Marta M Betcke, Matthias J Ehrhardt, and Carola-Bibiane Schönlieb. Gradient descent in a generalised Bregman distance framework. arXiv preprint arXiv:1612.02506, 2016.

M. Burger, E. Resmerita, and L. He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. Computing, 81(2-3):109–135, 2007.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. Found. Comput. Math., 9(6):717–772, 2009.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis., 40(1):120–145, 2011.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput., 20(1):33–61, 1998.

H. W. Engl, W. Heinz, M. Hanke, and A. Neubauer. Regularization of inverse problems, volume 375. Springer Science & Business Media, 1996.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In NeurIPS, pages 6151–6159, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. arXiv preprint arXiv:1802.08246, 2018.

B. Kaltenbacher, A. Neubauer, and O. Scherzer. Iterative regularization methods for nonlinear ill-posed problems, volume 6. Walter de Gruyter, 2008.

C. Molinari, M. Massias, L. Rosasco, and S. Villa. Iterative regularization for convex regularizers. In AISTATS, 2021.

J. Rasch and A. Chambolle. Inexact first-order primal–dual algorithms. Computational Optimization and Applications, 76(2):381–430, 2020.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. J. Mach. Learn. Res., 15(1):335–366, 2014.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Phys. D, 60(1-4):259–268, 1992.

R. Tibshirani. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol., 58(1):267–288, 1996.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. Constructive Approximation, 26(2):289–315, 2007.

Évaluation des risques liés aux pathogènes émis par l'irrigation de parcelles agricoles avec de l'eau usée traitée en station d'épuration à l'aide d'un réseau Bayésien

Gaspar Massiot[1], Dominique Courault[2] & Isabelle Albert[1]

[1] *UMR 518, Math-Info Appliquées, INRA-AgroParisTech 16, rue Claude Bernard, 75231 Paris Cedex 5, France*
[2] *UMR 1114 EMMAH, INRA, Université d'Avignon et des Pays du Vaucluse, Domaine St Paul, 84914 Avignon, France*

Résumé. Nous proposons d'adapter une approche Bayésienne développée pour l'évaluation des risques dans la chaine alimentaire à l'évaluation quantitative du risque microbien (QMRA) lié à la présence de pathogènes dans les bioaérosols. Nous construisons un modèle QMRA pour le risque lié à l'aérosolisation de *Legionella Pneumophila* lors de l'irrigation de parcelles agricoles par de l'eau usée traitée en station d'épuration puis par un pilote expérimental. Nous disposons de données de concentration en *Legionella Pneumophila* aux différentes étapes du traitement (entrée de station, sortie de station, sortie de pilote). Le modèle QMRA est augmenté par les données sous la forme d'un réseau Bayésien (BN). Un algorithme de Monte Carlo par chaîne de Markov (MCMC) permet enfin la mise à jour des quantités inconnues dans le modèle augmenté.

Mots-clés. *Legionella Pneumophila*, réseau Bayésien, inférence Bayésienne, MCMC, QMRA, réutilisation eau usée.

Abstract. Following recent development of Bayesian networks (BN) approaches in food safety risk assessment to study quantitative microbial risk assessment (QMRA) models, we propose to adapt this approach to the study of the risk associated to exposure to contaminated bioaerosols in the wastewater reuse field of research. We construct a QMRA model for the risks associated with the aerosolization of *Legionella Pneumophila* during the irrigation of agricultural parcels with wastewater sequencially treated in a treatment plant and in an experimental pilot. Experimental data are *Legionella Pneumophila* concentrations at different steps of the treatment (before entering treatment plant, at the exit of treatment plant, after going through the experimental pilot). We build the BN by linking the QMRA model to the data. We then ran an Markov chain Monte Carlo (MCMC) algorithm to update all the unknown quantities of the augmented model.

Keywords. *Legionella Pneumophila*, Bayesian network, Bayesian inference, MCMC, QMRA, wastewater reuse.

1 Introduction

Les modèles d'évaluation quantitative du risque microbien (QMRA) donnent un cadre rigoureux pour évaluer les risques de santé publique liés à la présence d'organismes pathogènes [6]. Dans le cadre de l'évaluation des risques liés à la présence de pathogènes dans l'air, ces modèles décrivent l'évolution du pathogène depuis son aérosolisation jusqu'à son inhalation par la population [7]. Plus particulièrement pour le développement d'un modèle QMRA pour les *Legionella Pneumophila*, la modélisation de l'exposition est souvent complexe car elle nécessite la prise en compte d'incertitudes dans la détection et l'identification des sources pathogènes de *Legionella*, de l'évolution du taux d'infectiosité et du transport spatial de la bactérie sous des conditions climatiques variables [11].

Les modèles QMRA sont souvent étudiés à l'aide d'approches par simulation de Monte Carlo (MC) qui consistent à tirer les valeurs des paramètres du modèles dans des distributions définies sur la base d'une connaissance externe (e.g., données expérimentales, historiques, dires d'experts) puis à propager ces valeurs initiales aux variables intermédiaires du système. Cette approche permet de construire des analyses de sensibilités et de prendre en compte la variabilité et l'incertitude des paramètres décrite dans les distributions sus-citées. En revanche, par définition, c'est une approche unidirectionnelle qui ne permet pas la mise à jour de ces distributions à l'aide de données recueillies en aval.

L'approche par réseaux Bayésiens (BN) permet la prise en compte directe de nouvelles données dans l'inférence. Les algorithmes de Monte Carlo par chaîne de Markov (MCMC) permettent son implémentation, y compris dans le cas de modèles complexes. Les développements récents autour des BN et leur application à des problématiques QMRA complexes pour la modélisation de l'évolution du risque tout au long de la chaîne alimentaire [1, 9, 10] ainsi que l'évaluation du risque lié à la présence de pathogènes dans l'eau [4, 5] permettent de penser que les BN présentent un fort potentiel méthodologique pour répondre aux problématiques QMRA [3], et plus particulièrement à notre problématique liée aux *Legionella*.

Ce travail se place dans le cadre du projet SmartFertiReuse (Smart Ferti-irrigation et RÉUtilisation des eaux USÉes traitées), financé par les Fonds Uniques Interministériels (FUI) et coordonné par la filière recherche et innovation de l'entreprise Véolia qui a pour objectif de développer un service complet pour accompagner le monde agricole et les collectivités locales dans une gestion agroécologique des eaux usées traitées et des fertilisants, depuis la conception d'un système opérationnel jusqu'au déploiement et pilotage à la parcelle en suivant la qualité de l'eau. Un site démonstrateur près de Tarbes, dans le sud-ouest de la France, a été retenu pour évaluer les différentes composantes agronomiques, économiques et sanitaires de cette pratique.

En nous basant sur une approche de Albert *et al.* [1], nous appliquons l'approche des BN à une modélisation QMRA pour l'évaluation des risques liés à l'aérosolisation de *Legionella Pneumophila* lors de l'irrigation par de l'eau usée.

TABLE 1 – Description des principales variables d'intérêt.

Module	Nœud	Unité	Description
Traitement des eaux usées	μ_X	CFU/L	Concentration en *Legionella* dans l'eau à l'étape $X \in \{0, 1, A\}$ du traitement, où 0 correspond à l'étape avant traitement par centrale de traitement des eaux (WWTP), 1 correspond à l'étape après traitement en WWTP, et A après traitement par pilote expérimental.
Aérosol	C^n	CFU/m^3	Concentration en *Legionella* dans l'air après irrigation en utilisant l'eau usée traitée à $n \in \{100, 300, 500, 1000\}$, mètres de la source
Exposition	D_C^n	$CFU/jour$	Dose moyenne d'exposition journalière pour les personnes de la catégorie $C \in \{F, R, P\}$ à n mètres de la source, où F correspond aux agriculteurs, R aux résidents et P aux passants.
Dose Réponse	$P_{inf,C}^n$	–	Probabilité d'infection par *Legionella* sur un jour d'exposition pour la catégorie C, à n mètres de la source.
	$P_{y,C}^n$	–	Probabilité d'infection par *Legionella* sur une année d'exposition pour la catégorie C, à n mètres de la source.
	$P_{life,C}^n$	–	Probabilité d'infection par *Legionella* sur 45 ans d'exposition pour la catégorie C, à n mètres de la source.

2 Modèle QMRA et Réseau Bayésien

Le modèle QMRA construit peut être découpé en 4 grands ensembles comme présenté dans la figure 1. Le tableau 1 présente une description des principales variables d'intérêt.

L'ensemble dit de *Contamination* correspond à la modélisation de la dispersion de la *Legionella Pneumophila* dans l'air, elle prend en compte les conditions climatiques à travers la vitesse du vent et l'ensoleillement (u, E) ainsi que la quantité de *Legionella Pneumophila* aérosolisée Q comme décrit dans l'équation (1). Un modèle de dispersion atmosphérique à panache gaussien permet d'estimer la concentration en *Legionella Pneumophila* à différentes distances de la source.

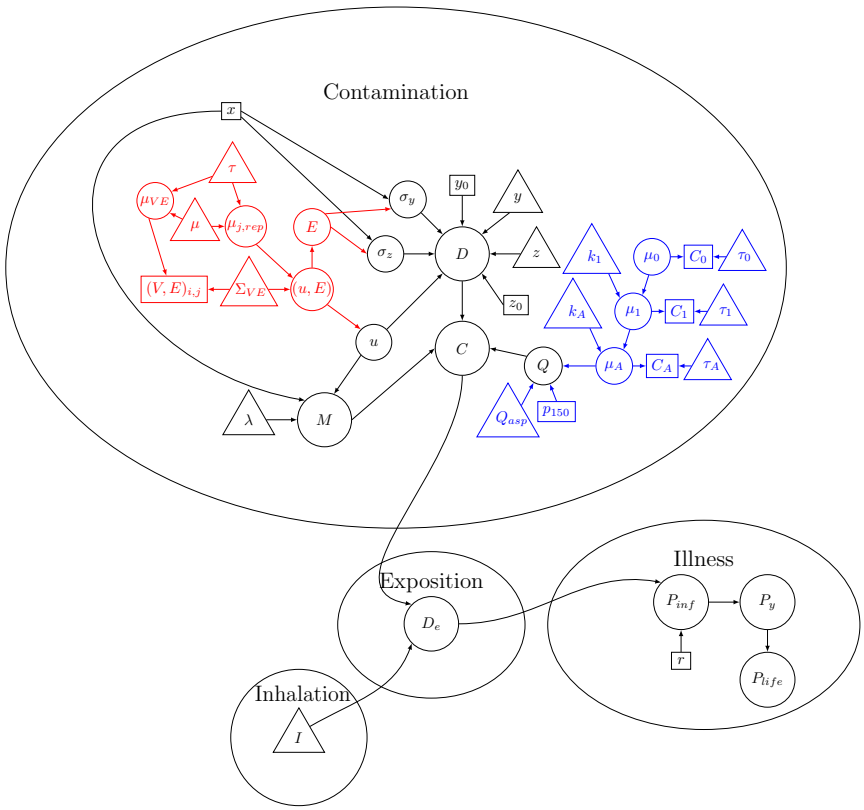


FIGURE 1 – Représentation graphique du BN mis en place pour la modélisation. Les triangles représentent les paramètres aléatoires, les rectangles représentent les données ou les nœuds fixés et les cercles représentent les variables intermédiaires. Le graphe noir correspond au modèle QMRA, les variables ajoutées en bleu et en rouge correspondent à l'augmentation du modèle en un réseau bayésien. Les principales variables d'intérêt sont décrites dans le tableau 1.

TABLE 2 – Catégories de population exposée

Label	**Activité**	**Temps d'exposition**
Agriculteur	travail sur les cultures pendant et après l'aspersion (surveillance, entretien, récolte, etc.)	2h par jour d'irrigation
Résident	résidence à proximité des terrains irrigués	2 secondes toutes les minutes pendant les 2.27h passées dans le jardin
Passant	passage à proximité des terrains irrigués pendant l'aspersion	1 minute par jour

$$C^n = \frac{Q}{2\pi u \sigma_y \sigma_z} \exp\left[-\frac{y^2}{2\sigma_y^2}\right] \left\{ \exp\left[-\frac{(z-H_e)^2}{2\sigma_z^2}\right] + \exp\left[-\frac{(z+H_e)^2}{2\sigma_z^2}\right] \right\} \exp^{-\frac{\lambda n}{u}}, \quad (1)$$

où C^n correspond à la concentration en *Legionella* dans l'air à $n \in \{100, 300, 500, 1000\}$ mètres de la buse d'aspersion (nombre par m^3), y représente la distance horizontale orthogonale à la direction du vent (m), z représente la hauteur du récepteur $(1.5m)$, Q représente le taux d'émission du pathogène par la buse d'aspersion (nombre par s), H_e représente la hauteur de la source (m), u représente la vitesse du vent (m/s), λ est un coefficient d'inactivation du pathogène [8] (s^{-1}), enfin σ_y et σ_z sont respectivement les coefficients de dispersion horizontale et verticale estimés à partir des données météorologiques grâce aux classes de Pasquill [8].

Le taux d'exposition est estimé pour trois catégories de population présentées dans le tableau 2 inspiré par la méthodologie mise en place dans le rapport d'expertise collective de l'Anses (Saisine n° 2009-SA-0329) pour l'évaluation des risques sanitaires liés à la réutilisation des eaux usées traitées pour l'irrigation par aspersion des cultures et des espaces verts, ainsi que pour le lavage des voiries.

Enfin, nous utilisons le modèle dose-réponse proposé par Armstrong and Haas (2007) [2] pour la légionellose :

$$P_{inf} = 1 - e^{-rD},$$

où r représente le taux d'infectivité par unité de *L. pneumophila*, D représente la dose de *L. pneumophila* et P_{inf} représente la probabilité d'être infecté après inhalation de la dose D de *L. pneumophila*.

Les données (vent, ensoleillement et concentration en *L. Pneumophila* dans l'eau) sont prises en compte dans les sous-graphes rouge et bleu de la figure 1.

Enfin, un algorithme MCMC a été implémenté pour mettre à jour les connaissances a priori sur toutes les quantités présentes dans le BN.

Références

[1] I. Albert, E. Grenier, J.-B. Denis, and J. Rousseau. Quantitative risk assessment from farm to fork and beyond : A global bayesian approach concerning food-borne diseases. *Risk Analysis : An International Journal*, 28(2) :557–571, 2008.

[2] T. W. Armstrong and C. N. Haas. Quantitative microbial risk assessment model for Legionnaires' disease : assessment of human exposures for selected spa outbreaks.

[3] D. Beaudequin, F. Harden, A. Roiko, H. Stratton, C. Lemckert, and K. Mengersen. Beyond qmra : Modelling microbial health risk as a complex system using bayesian networks. *Environment international*, 80 :8–18, 2015.

[4] R. Goulding, N. Jayasuriya, and E. Horan. A bayesian network model to assess the public health risk associated with wet weather sewer overflows discharging into waterways. *water research*, 46(16) :4933–4940, 2012.

[5] A. D. Gronewold, C. A. Stow, K. Vijayavel, M. A. Moynihan, and D. R. Kashian. Differentiating enterococcus concentration spatial, temporal, and analytical variability in recreational waters. *Water research*, 47(7) :2141–2152, 2013.

[6] C. N. Haas, J. B. Rose, and C. P. Gerba. *Quantitative microbial risk assessment*. John Wiley & Sons, 1999.

[7] K. A. Hamilton and C. N. Haas. Critical review of mathematical approaches for quantitative microbial risk assessment (QMRA) of Legionella in engineered water systems.

[8] K. A. Hamilton, M. T. Hamilton, W. Johnson, P. Jjemba, Z. Bukhari, M. LeChevallier, and C. N. Haas. Health risks from exposure to legionella in reclaimed water aerosols : Toilet flushing, spray irrigation, and cooling towers. *Water Research*, 134 :261–279, 2018.

[9] C. S. Rigaux Anselet, F. Carlin, C. Nguyen-thé, and I. Albert. Inferring an augmented bayesian network to confront a complex quantitative microbial risk assessment model with durability studies : application to bacillus cereus on a courgette purée production chain. *Risk analysis*, 33(5) :877–892, 2013.

[10] J. Smid, L. Heres, A. Havelaar, and A. Piclaat. A biotracing model of salmonella in the pork production chain. *Journal of food protection*, 75(2) :270–280, 2012.

[11] H. Whiley, A. Keegan, H. Fallowfield, and K. Ross. Uncertainties associated with assessing the public health risk from legionella. *Frontiers in microbiology*, 5 :501, 2014.

On reparameterisations of the Poisson process model for extremes in a Bayesian framework

Théo Moins[1] & Julyan Arbel[1] & Anne Dutfoy[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
{theo.moins, julyan.arbel, stephane.girard}@inria.fr
[2] *EDF R&D dept. Péricles, 91120 Palaiseau, France*
anne.dutfoy@edf.fr

Résumé. Combiner l'analyse des valeurs extrêmes avec des méthodes bayésiennes a plusieurs avantages, comme la prise en compte d'information *a priori* ou encore la possibilité d'étudier des cas irréguliers en statistique fréquentiste. Nous nous attardons ici sur un modèle d'extrêmes par processus de Poisson, et proposons une approche alternative à une étude récente sur une reparamétrisation du modèle qui orthogonalise les paramètres pour améliorer l'échantillonnage *a posteriori* par méthode de Monte-Carlo par chaînes de Markov (MCMC).

Mots-clés. Théorie des valeurs extrêmes, Processus de Poisson, Inférence bayésienne.

Abstract. Combining extreme value analysis with Bayesian methods has several advantages, such as the consideration of prior information or the ability to study irregular cases for frequentist statistics. We focus here on a model of extremes by Poisson process, and propose an alternative of a recent study on a parameterisation of the model which orthogonalizes the parameters to improve posterior sampling by Markov chain Monte-Carlo method (MCMC).

Keywords. Extreme-Value theory, Poisson processes, Bayesian inference.

1 Introduction

Bayesian inference provides tools to estimate parameters of extreme value models, quantify their uncertainty (Arbel et al., 2019), and exploit prior information if available. In particular, Markov chain Monte Carlo (MCMC) methods have various advantages for extreme value parameter inference, which are summed up in Coles and Tawn (1996). However, the interdependence of the parameters can compromise their estimation, and in particular the convergence of the Markov chain. Among the extreme models, the Poisson process allows generalising the two most frequent models, namely block maxima and peak-over-threshold. After a presentation of the extreme value model based on a Poisson process, we present two reparametrisations to facilitate Bayesian inference. Both approaches are then compared from a theoretical and experimental point of view.

2 Poisson process characterisation of extremes

Let $(X_1, \ldots, X_n)$ be i.i.d. random variables with distribution function F, and $M_n = \max\{X_1, \ldots, X_n\}$, whose distribution is F^n. The domain of attraction of F is defined to be the set of distributions G such that there exist two sequences $a_n > 0$ and b_n such that

$$F^n(a_n x + b_n) \to G(x) \text{ as } n \to \infty.$$

The extreme-value theorem states that if F belongs to the maximum domain of attraction of a distribution G, then G is a generalised extreme value (GEV) distribution:

$$G(x) = G(x \mid \boldsymbol{\theta}) = \begin{cases} \exp\left(-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-\frac{x-\mu}{\sigma})) & \text{if } \xi = 0, \end{cases}$$

where $\{x\}_+ = \max\{0, x\}$ and $\boldsymbol{\theta} = (\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$. Asymptotically, a_n and b_n can be absorbed in the parameters in such a way that it is sufficient to estimate $\boldsymbol{\theta}$.

An alternative is to consider values that exceed a high threshold u. A second extreme-value theorem, known as the Pickands theorem, states that if F belongs to the maximum domain of attraction of $G(\cdot \mid \mu, \sigma, \xi)$, then the distribution function of the exceedances $P(X - u \mid X > u)$ can be approximated by a Generalised Pareto Distribution (GPD):

$$P(X < y + u \mid X > u) \xrightarrow[u \to x^*]{} \begin{cases} 1 - \left\{1 + \xi\frac{y}{\tilde{\sigma}_u}\right\}_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ 1 - \exp\left(-\frac{y}{\tilde{\sigma}_u}\right) & \text{if } \xi = 0, \end{cases}$$

where x^* is the upper endpoint of X. The shape parameter ξ is the same as in the GEV model, and the relation $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$ links the two scale parameters.

Summarised by Coles (2001), a third way to characterise extreme observations comes from the theory of point processes, and unifies the two previous models. From the extreme value theorem, one can show that the point process N_n related to the point sequence $\{X_1, \ldots, X_n\}$ is such that for intervals $I_u = [u, +\infty)$ with sufficiently large u, we have

$$N_n(I_u) \sim \mathcal{B}(n, p)\ \text{, with } p \approx \begin{cases} \frac{1}{n}\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \frac{1}{n}\exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases}$$

As $n \to +\infty$, the binomial distribution converges to a Poisson distribution $\mathcal{P}(\Lambda(I_u))$, with

$$\Lambda(I_u) = \begin{cases} \left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \tag{1}$$

This result combined with the independence property for distributions of N_n on non-overlapping sets is sufficient to say that N_n converges to a non-homogeneous Poisson process, with intensity measure for a fixed u given by (1):

$$N_n \xrightarrow{d} N, \quad \text{with } N(I_u) \sim \mathcal{P}(\Lambda(I_u)).$$

Clearly the GEV model is a special case of this limiting Poisson process, since:

$$P(M_n < z) = P(N_n(I_z) = 0) \xrightarrow[n\to+\infty]{} P(N(I_z) = 0) = \exp(-\Lambda(I_z)) = G(z \mid \boldsymbol{\theta}).$$

However, the parameters (μ, σ, ξ) are here related to the overall maximum of the dataset, and it is frequent to study maxima of m smaller blocks (such as annual maxima). To this end, the intensity measure $\Lambda(I_u)$ is multiplied by a scaling factor m equal to the number of blocks. In the same way, the threshold excess model is also a special case as it can be derived from the Poisson process model. One difference between estimation with GPD and Poisson process is the threshold dependence of $\tilde{\sigma}_u$ on the first case, contrary to σ which is here the same as in the GEV model, and does not depend on u.

Then, we are able to define the likelihood for n observations $(x_1, \ldots, x_n)$ above u:

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = \exp\left(-m\Lambda(I_u \mid \boldsymbol{\theta})\right) \prod_{i=1}^{n} \lambda(x_i \mid \boldsymbol{\theta}),$$

$$\text{with} \quad \lambda(x \mid \boldsymbol{\theta}) = \begin{cases} \sigma^{-1}\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1+\xi}{\xi}} & \text{if } \xi \neq 0\,, \\ \sigma^{-1}\exp(-\frac{x-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases}$$

The choice of the scaling factor m will have an impact on μ and σ, but ξ stays invariant with respect to m. More precisely, Wadsworth et al. (2010) show that if (μ_1, σ_1, ξ) are parameters associated with m_1 block maxima and (μ_2, σ_2, ξ) are parameters associated with m_2 block maxima, then

$$\mu_2 = \mu_1 - \frac{\sigma_1}{\xi}\left(1 - \left(\frac{m_2}{m_1}\right)^{-\xi}\right), \quad \text{and} \quad \sigma_2 = \sigma_1\left(\frac{m_2}{m_1}\right)^{-\xi}. \tag{2}$$

3 Two reparameterisations for Bayesian inference

Parameter orthogonality is defined as a property of a set of parameters $\boldsymbol{\theta}$ which leads to a diagonal Fisher information matrix $I(\boldsymbol{\theta})$. Sharkey and Tawn (2017) show empirically that orthogonality improves the convergence of the MCMC to the joint posterior distribution of the parameters. We successively describe their method to achieve near-orthogonality, and suggest another one proposed by Chavez-Demoulin and Davison (2005) for another purpose, that seems to be more effective for this objective.

3.1 Near-orthogonality by tuning m (Sharkey and Tawn, 2017)

In view of (2), it is easy to transform the parameters μ and σ by changing the value of the scaling factor. Sharkey and Tawn (2017) use this degree of freedom to find an m that minimises the asymptotic covariance between parameters, in order to change it

before using the Metropolis–Hastings algorithm, and at the end, restore the parameters corresponding to the initial number of blocks. As the asymptotic covariance matrix can be found by inverting the Fisher information matrix, the authors describe their problem as finding a value of m that near-orthogonalizes the parameters (μ, σ, ξ).

The details of the derivation will not be given here, but one should note that they are only valid for $\xi > -1/2$. In the end, asymptotic covariances can be written as functions of $\left(x = -\frac{1}{\xi}\log\left\{1+\xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+, \sigma, \xi\right)$ as:

$$\begin{aligned}
\mathrm{ACov}(\mu,\sigma) &= \frac{\sigma^2}{m\xi^2}e^x\left(\xi(1+\xi)^2x^2-(1+3\xi)(1+\xi)x+\xi^3+(1+\xi)(1+2\xi)\right.\\
&\quad \left.+e^{-\xi x}(1+\xi)(1+2\xi)(x-1)\right),\\
\mathrm{ACov}(\mu,\xi) &= \frac{\sigma}{m\xi^2}e^x(1+\xi)\left(\xi(1+\xi)x-(1+2\xi)(1-e^{-\xi x})\right),\\
\mathrm{ACov}(\sigma,\xi) &= \frac{\sigma}{m}e^x(1+\xi)\left((1+\xi)x-1\right).
\end{aligned}$$

Denoting by ρ_{θ_1,θ_2} the asymptotic correlation between two of the three parameters, the authors noted that a range of values can also work for m between m_1 and m_2, where

$$m_1 = \underset{m}{\mathrm{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\mu,\xi}|\} \text{ and } m_2 = \underset{m}{\mathrm{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\sigma,\xi}|\}.$$

They also found on their experiments that m_1 cancels $\mathrm{ACov}(\mu,\sigma)$, and that m_2 cancels $\mathrm{ACov}(\sigma,\xi)$. A numerical method is used in Sharkey and Tawn (2017) to approximate m_1 and m_2 as functions of ξ, so their framework consists of estimating ξ (with maximum likelihood for example), to obtain $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$ and choose a value in this interval to run the MCMC.

3.2 Orthogonal parameterisation (our proposal)

More directly, there exists a parameterisation of the Poisson process that leads directly to orthogonality. Suggested by Chavez-Demoulin and Davison (2005), it consists of the following change of variable:

$$(r,\nu,\xi) = \left(m\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}, (1+\xi)(\sigma+\xi(u-\mu)), \xi\right).$$

Parameter r represents the intensity of the Poisson process, which is the expected number of exceedances, and the two others can be seen as an orthogonal parametrisation of the GPD distribution with scale $\tilde{\sigma}_u = \sigma + \xi(u-\mu)$ and shape ξ. The motivations of Chavez-Demoulin and Davison (2005) for this transformation are different[1], but this parameterisation can be used here to ensure full orthogonality, which means that $I(r,\nu,\xi)$ is diagonal.

[1]To avoid computations difficulties when using a generalised linear model for extremes.

4 Comparison of the two approaches

The approach of Sharkey and Tawn (2017) implies to study the x roots of $\text{ACov}(\mu, \sigma)$ and $\text{ACov}(\sigma, \xi)$ to respectively deduce $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. The corresponding value of m can be found by observing that $x = \log(\frac{r}{m})$, with $r = m\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}$ the expected number of observations. Although r is unknown before estimation, it can be easily estimated by the actual number of observations n, allowing us to deduce the value of m from a given x. The authors provide an approximation method to numerically compute $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. In contrast here, we investigate the existence, uniqueness and position of the roots from a theoretical point of view.

For $\text{ACov}(\sigma, \xi)$, we directly have $x_1 = \frac{1}{1+\xi}$ as the unique root. Moreover, as $\xi > -\frac{1}{2}$, we have $x_1 > 0$, which motivates us to study the sign of the root x_2 for $\text{ACov}(\mu, \sigma)$. Indeed, if x_2 is unique and $x_2 < 0$, then the choice $x = 0$ which cancels the third asymptotic covariance $\text{ACov}(\mu, \xi)$ will always be reasonable as it will stay in the targeted interval, between the two other roots. In addition, $x = 0$ corresponds to the choice $m = r$ (which in practice translates into $m = n$), and is a simple choice as it does not require any estimation of ξ. The interest of the choice $m = n$ has already been mentioned in Wadsworth et al. (2010) to improve the mixing property of the chain. Unfortunately, a study of function for $\text{ACov}(\mu, \sigma)$ shows that the properties of uniqueness and positivity are only valid in the case where $\xi > 0$. In that case, studies of Wadsworth et al. (2010) and Sharkey and Tawn (2017) corroborate the choice of $m = n$. However, it is not the case when $-\frac{1}{2} < \xi < 0$. The study shows that x_2 is not negative here, and worse, may not be unique (Fig. 1).

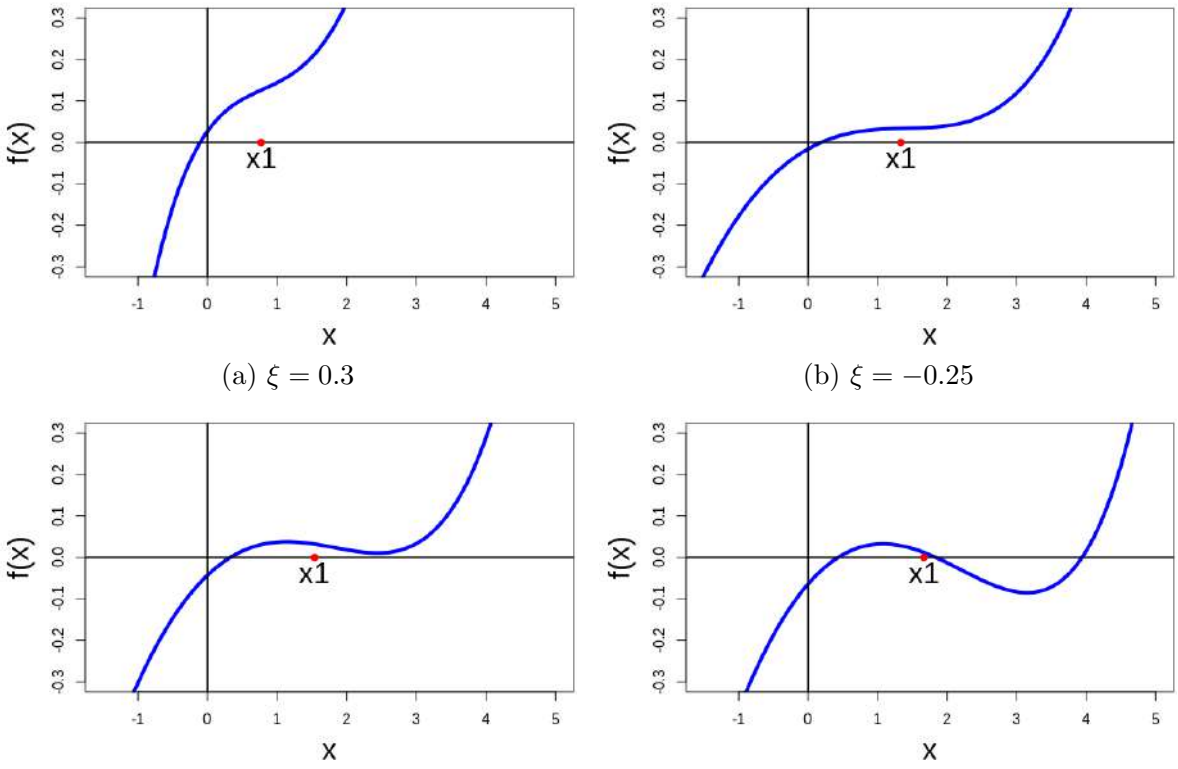
Thus, instead of tuning m, using the orthogonal parameterisation of Chavez-Demoulin and Davison (2005) is more adapted, as has the following three advantages:

1. It exactly orthogonalises the three parameters.
2. It is more accurate in the sense that there is no need to estimate r (as r is one of the parameters).
3. It is valid for all $\xi > -1/2$.

Moreover, by plugging the variables (r, ν) in (2), we can show that the invariance property with respect to m holds for the three parameters, and so the parametrisation is totally independent from the choice of m. In the communication, we shall also present the experimental part, comparing the convergence and mixing properties of the Markov chains corresponding to different parameterisations on various datasets.

References

Arbel, J., Crispino, M., and Girard, S. (2019). "Dependence properties and Bayesian inference for asymmetric multivariate copulas." *Journal of Multivariate Analysis*, 174.



(a) $\xi = 0.3$ (b) $\xi = -0.25$

(c) $\xi = -0.35$ (d) $\xi = -0.4$

Figure 1: Graphs of $\text{ACov}(\mu, \sigma)$ as a function of x for different values of ξ. When $\xi > 0$ (a), 0 effectively belongs to the segment $[x_1, x_2]$, but it is not the case anymore when $\xi < 0$, with different scenarios. The covariance function can stay monotonic (b), or may be locally decreasing (c), up to having multiple roots (d).

Chavez-Demoulin, V. and Davison, A. C. (2005). "Generalized additive modelling of sample extremes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 207–222.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. London: Springer-Verlag.

Coles, S. G. and Tawn, J. A. (1996). "A Bayesian analysis of extreme rainfall data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4), 463–478.

Sharkey, P. and Tawn, J. A. (2017). "A Poisson process reparameterisation for Bayesian inference for extremes." *Extremes*, 20(2), 239–263.

Wadsworth, J., Tawn, J. A., and Jonathan, P. (2010). "Accounting for choice of measurement scale in extreme value modeling." *The Annals of Applied Statistics*, 4(3), 1558–1578.

Comparaison de multiples critères de performance de prédictions dynamiques

Clémence Moreau [1] & Jérémie Riou [2,3] & Marine Roux [1]

[1] *HIFIH, UPRES 3859, SFR 4208, Angers University, France (presenting author)*
cl.moreau@univ-angers.fr ; marine.roux@univ-angers.fr
[2] *MINT, UMR INSERM 1066, CNRS 6021, Angers University, France*
[3] *Methodology and Biostatistics Department, Delegation to Clinical Research and Innovation, Angers University Hospital, France ; jeremie.riou@univ-angers.fr*

Résumé. Des versions dynamiques de l'aire sous la courbe ROC (AUC) et du Brier score ont été développées par Blanche *et al.* [2] dans le cadre des risques compétitifs afin de quantifier les capacités prédictives des modèles conjoints. Certains tests ont également été proposés afin de les comparer. Cependant, seules deux prédictions peuvent être comparées. Le résultat principal de notre étude est la proposition d'une nouvelle procédure de test permettant d'étendre cette comparaison à plus de deux prédictions.

Mots-clés. prédiction dynamique, AUC & Brier score dynamiques, analyse de survie

Abstract. Dynamic versions of the area under the ROC curve (AUC) and the Brier score have been developed by Blanche *et al.* [2] in the competing risks setting to quantify the predictive accuracy of the joint models. Some tests have also been proposed for comparison. However, only two predictions can be compared. The main result of this study is a new test procedure to extend this comparison to more than two predictions.

Keywords. dynamic prediction, dynamic AUC & Brier score, survival analysis

1 Introduction

Un des enjeux majeurs de la pratique clinique actuelle est l'identification des marqueurs les plus aptes à prédire les risques individuels liés à une maladie. En effet, cela permet aux cliniciens d'agir précocement, limitant ainsi la morbidité des patients vulnérables. Le développement de la modélisation dite dynamique contribue à cette fin. Elle permet d'établir des prédictions de risques individuelles dynamiques au sens où elles sont réactualisées au fil du temps dès que le profil clinique du patient s'étoffe. En particulier, les modèles conjoints [5, 6] présentent l'avantage d'établir leurs prédictions en tenant compte de tout l'historique du patient. Avant d'utiliser ces modèles en pratique clinique, leurs capacités prédictives doivent être évaluées rigoureusement.

Pour ce faire, des versions dynamiques de l'aire sous la courbe ROC (AUC) et du Brier Score (BS) ont été développées par Blanche *et al.* [2] dans le cadre des risques compétitifs.

Un test de comparaison entre deux prédictions est également établi. Or, grâce aux récents progrès technologiques, de plus en plus de marqueurs sont disponibles, induisant le besoin de comparaison de multiples marqueurs. Notre travail permet de répondre à cette problématique.

Application clinique motivant la recherche. Aussi appelée la maladie du foie gras, la *Non Alcoholic Fatty Liver Disease* (NAFLD) touche principalement des personnes atteintes d'obésité et peut mener à de graves complications hépatiques. Parmi toutes les lésions hépatiques observées dans les cas de NAFLD, la fibrose hépatique est le principal indicateur du pronostic du patient. La biopsie du foie est la référence pour l'évaluation de ces lésions, mais cette procédure ne peut être largement utilisée dans la vaste population de la NAFLD en raison de son caractère invasif et des coûts associés. Des tests non-invasifs sont maintenant disponibles pour l'examen de la fibrose hépatique. Sans risque pour le patient, ce sont principalement des tests sanguins et des mesures de l'élasticité du foie pouvant facilement être répétés lors du suivi du patient. Nous utiliserons notre procédure de test afin d'identifier le meilleur marqueur pronostique parmi les tests non-invasifs disponibles pour les complications hépatiques des patients atteints de NAFLD.

2 Méthodes

2.1 Notations

Soit T la variable de temps à l'évènement, C le temps de censure, et $\Delta = \mathbb{1}(T \leqslant C)$ l'indicateur d'évènement. Par souci de simplification, seulement deux évènements compétitifs sont pris en compte : $\eta = 1$ pour l'évènement principal, $\eta = 2$ pour le compétitif. Soit M un marqueur pour lequel $\mathcal{M}_i(s)$ représente l'information collectée pour l'individu i jusqu'au temps landmark s. Nous considérons que la loi jointe de $(T, \eta, \mathcal{M}(s))$ est approchée par un modèle conjoint paramétré par un vecteur de paramètres ξ, dont l'estimateur est $\widehat{\xi}$.

Soit $h \in \mathbb{R}^+$ un horizon de prédiction fixé. La prédiction dynamique $\pi_i(s,h)$ [4] correspond à la probabilité pour un individu i de subir l'évènement principal dans l'intervalle de temps $]s, s+h]$ étant donnée l'information collectée jusqu'au temps s. Elle est définie par

$$\pi_i(s,h) = \mathbb{P}_{\widehat{\xi}}(s < T_i \leqslant s+h, \eta_i = 1 \mid T_i > s, \mathcal{M}_i(s), X_i), \tag{1}$$

avec $\mathbb{P}_{\widehat{\xi}}$ la distribution de probabilité paramétrée par $\widehat{\xi}$, et X_i un vecteur de variables à baseline. De plus, soient $\widetilde{T} = \min(T, C)$ le temps d'observation, et $\widetilde{\eta} = \Delta\eta$ où $\widetilde{\eta}_i = 1$ si l'individu i a subi l'évènement principal (2 si évènement concurrent) avant d'être censuré et 0 sinon.

2.2 Critères de performance des prédictions dynamiques

Soit un n-échantillon $\left\{\left(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \pi_i(\cdot,\cdot)\right), i = 1, \ldots, n\right\}$. Dans cette partie, nous nous intéressons à la prédiction de l'évènement principal ($\eta = 1$).

AUC dynamique. L'AUC dynamique est utilisé afin de mesurer la discrimination de prédictions dynamiques. Soit $D_i(s,h) = \mathbb{1}_{(s<T_i\leqslant s+h, \eta_i=1)}$ égal à 1 si l'individu i subit l'évènement principal entre $]s, s+h]$, 0 sinon. L'AUC dynamique de Blanche *et al.* [2] au temps landmark s et à l'horizon h est défini par, pour tout $i, j \in \{1, \ldots, n\}$,

$$AUC(s,h) = \mathbb{P}\left(\pi_i(s,h) > \pi_j(s,h) \mid D_i(s,h) = 1, D_j(s,h) = 0, T_i > s, T_j > s\right).$$

Afin de prendre en compte la censure lors de l'estimation, Uno *et al.* [9] et Hung & Chiang [3] ont développé des estimateurs *Inverse Probability of Censoring Weighting* (IPCW). Ce type d'estimateur consiste à pondérer les individus par la probabilité de ne pas être censuré. En supposant le temps de censure C indépendant de $(T, \eta, \pi(\cdot,\cdot))$, les poids sont définis de la façon suivante

$$\widehat{W}_i(s,h) = \frac{\mathbb{1}_{\left(\widetilde{T}_i > s+h\right)}}{\widehat{G}(s+h \mid s)} + \frac{\mathbb{1}_{\left(s<\widetilde{T}_i\leqslant s+h\right)}\Delta_i}{\widehat{G}\left(\widetilde{T}_i \mid s\right)},$$

où $\widehat{G}(u)$ est l'estimateur de Kaplan-Meier de $\mathbb{P}(C > u)$, et $\widehat{G}(u \mid s) = \widehat{G}(u)/\widehat{G}(s)$ la probalilité de ne pas être censuré à u sachant qu'on ne l'est pas à s. Soit $\widetilde{D}_i(s,h) = \mathbb{1}_{\left(s<\tilde{T}_i\leqslant s+h, \tilde{\eta}_i=1\right)}$. Ainsi, Blanche *et al.* [2] proposent d'estimer $AUC(s,h)$ par

$$\widehat{AUC}(s,h) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{(\pi_i(s,h)>\pi_j(s,h))}\widetilde{D}_i(s,h)\left(1-\widetilde{D}_j(s,h)\right)\widehat{W}_i(s,h)\widehat{W}_j(s,h)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{D}_i(s,h)\left(1-\widetilde{D}_j(s,h)\right)\widehat{W}_i(s,h)\widehat{W}_j(s,h)}. \quad (2)$$

Brier score dynamique. Afin de mesurer la discrimination et la calibration de prédictions dynamiques, Blanche *et al.* [2] ont également développé la définition du Brier score dynamique

$$BS(s,h) = \mathbb{E}\left[\left(D(s,h) - \pi(s,h)\right)^2 \mid T > s\right]$$

grâce notamment à la définition du Brier score à risque compétitif de Schoop *et al.* [7]. Il est estimé par

$$\widehat{BS}(s,h) = \frac{1}{n\widehat{S}_{\tilde{T}}(s)}\sum_{i=1}^{n}\widehat{W}_i(s,h)\left(\tilde{D}_i(s,h) - \pi_i(s,h)\right)^2,$$

où $\widehat{S}_{\tilde{T}}(s)$ est la probabilité estimée d'observer un individu à risque au temps s.

2.3 Procédure de comparaison

Pour pouvoir comparer plus de deux prédictions dynamiques, nous faisons dans un premier temps une comparaison globale afin de voir si l'une d'entre elles diffère des autres. Si oui, nous effectuons ensuite des tests post-hoc afin de localiser cette différence. La théorie de ce test étant la même pour les versions dynamiques de l'AUC et du Brier score, nous posons $\Theta(\cdot, h)$ désignant aussi bien $AUC(\cdot, h)$ que $BS(\cdot, h)$ à l'horizon h, et $\widehat{\Theta}(\cdot, h)$ l'estimateur IPCW associé. Soient Θ_p, $p \in \mathcal{P} = \{1, \ldots, P\}$, les p critères de qualité ici comparés, disponibles au temps landmark $s \in \mathcal{S} = \{1, \ldots, S\}$.

Procédure.

- **Étape 1 : Test global.** Nous testons

$$\mathcal{H}_0 : \Theta_1(\cdot, h) = \ldots = \Theta_P(\cdot, h)$$

versus

$$\mathcal{H}_1 : \exists p \in \mathcal{P}, \Theta_p(\cdot, h) \neq \Theta_1(\cdot, h)$$

où Θ_1 est le critère de qualité associé au marqueur de référence posé arbitrairement, via la statistique de test

$$\Psi_{\mathcal{S},\mathcal{P}}(h) = \sum_{s=1}^{S} \sum_{p=2}^{P} \left[\frac{\sqrt{n}}{\widehat{\sigma}^{p,1}_{s,h}} \left(\widehat{\Theta}_p(s, h) - \widehat{\Theta}_1(s, h) \right) \right]^2 ,$$

Sous $\mathcal{H}_0$,

$$\Psi_{\mathcal{S},\mathcal{P}}(h) \xrightarrow{\mathcal{L}} \Gamma\left(\frac{P-1}{z}; z \right), \text{ où } z = 2 \left(1 + \frac{2(P-1) \sum_{i \neq j}^{S} \rho_{ij}}{S(P-1)} \right) ;$$

avec $\xrightarrow{\mathcal{L}}$ désignant la convergence en loi et $\rho_{ij} = Corr\left(\Theta(i, h), \Theta(j, h)\right)$ la corrélation entre les différents temps landmark.

- **Étape 2 : Test post-hoc.** Si le test global montre une différence significative lors de la première étape, nous effectuons toutes les comparaisons par paires afin de situer cette différence.

Gestion de la multiplicité. La gestion de la multiplicité des tests permettant un contrôle du *Family-Wise Error Rate* (FWER) au seuil 5% se fera à l'aide la procédure de Shaffer [8], montrant d'excellentes performances dans ce contexte [1]. La procédure de Shaffer étend celle de Bonferroni-Holm dans le cadre des dépendances logiques entre les hypothèses.

3 Simulations

Des premières études de simulation ont été réalisées pour le test global. Dans un premier temps, nous comparons notre test global au test simultané déjà existant de Blanche *et al.* [2] dans le cadre de comparaison de deux prédictions dynamiques. Nous avons également étudié l'erreur de type I et la puissance obtenue lorsque plus de deux prédictions sont comparées. Les résultats sont résumés dans le table 1.

Simulations	**Individus**	**Hypothèse**	ϕ [a]	φ [b]	ψ [c]
5 000	500	$\mathcal{H}_0$	5.68	3.34	4.36
5 000	500	$\mathcal{H}_1$	43.24	36.02	49.82
1 000	500	$\mathcal{H}_0$	6.1	3.8	-
1 000	500	$\mathcal{H}_1$	42.8	34.6	-
1 000	1 000	$\mathcal{H}_0$	4.6	3.2	-
1 000	1 000	$\mathcal{H}_1$	76.0	68.4	-

a. Résultats correspondant au test simultané de Blanche *et al.* [2].
b. Résultats correspondant à notre test de comparaison global pour deux prédictions dynamiques.
c. Résultats correspondant à notre test de comparaison global pour trois prédictions dynamiques.

TABLE 1 – Résultats des études de simulations pour le test global sur l'ensemble des temps landmark $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ et l'horizon de prédiction $h = 5$ ans.

Dans le cadre d'une comparaison de deux prédictions dynamiques, le test global permet un contrôle de l'erreur de type I dans l'ensemble des scénarios envisagés tout en obtenant une puissance proche du test de Blanche *et al.* [2]. Dans le cadre de trois prédictions dynamiques, l'erreur de type I est également contrôlée.

Conclusion. Dans le cadre de la NAFLD, notre nouvelle procédure permet de sélectionner le meilleur biomarqueur parmi une grande quantité disponible afin de prédire les complications hépatiques. D'après les premiers résultats de simulations, l'erreur de type I est contrôlée dans tous les scénarios considérés pour notre test global. La procédure de Shaffer a permis par la suite de contrôler le FWER lors des comparaisons par paires. Finalement, ce travail permet de comparer plus de deux critères de performance de prédictions dynamiques et sera disponible par le biais d'un package R prochainement.

Références

[1] Paul Blanche, Jean-François Dartigues, and Jérémie Riou. A closed max-t test for multiple comparisons of areas under the roc curve. *Biometrics*, 2020.

[2] Paul Blanche, Cécile Proust-Lima, Lucie Loubère, Claudine Berr, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1) :102–113, 2015.

[3] Hung Hung and Chin-Tsang Chiang. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1) :8–26, 2010.

[4] Cécile Proust-Lima and Paul Blanche. Dynamic predictions. *Wiley StatsRef : Statistics Reference Online*, pages 1–6, 2014.

[5] Cécile Proust-Lima and Jeremy MG Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa : a joint modeling approach. *Biostatistics*, 10(3) :535–549, 2009.

[6] Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3) :819–829, 2011.

[7] Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1) :88–112, 2011.

[8] Juliet Popper Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395) :826–831, 1986.

[9] Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478) :527–537, 2007.

Explicitly estimating shifts in species' optimum position along environmental gradients

Bastien Mourguiart [1] , Benoit Liquet [2] , Thibaut Couturier [3] , Jérôme Mansons [4] , Yoan Braud [5] , Kerrie Mengersen [6] & Aurélien Besnard [7]

[1] *CNRS/Univ Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications de Pau - Fédération MIRA, UMR 5142, 64600 Anglet, France, bastien.mourguiart@univ-pau.fr*

[2] *CNRS/Univ Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications de Pau - Fédération MIRA, UMR 5142, 64600 Anglet, France, benoit.liquet@univ-pau.fr*

[3] *Centre d'Ecologie Fonctionnelle et Evolutive - UMR 5175 - CNRS, EPHE-PSL University, Université Paul Valéry Montpellier 3, Montpellier, 34293, France, Thibaut.COUTURIER@cefe.cnrs.fr*

[4] *Parc National du Mercantour, Nice, 06006, France, jerome.mansons@mercantour-parcnational.fr*

[5] *Bureau d'études Entomia, Vaumeilh, 04200, France, yoan_braud@yahoo.fr*

[6] *ARC Centre of Excellence in Mathematical and Statistical Frontiers, School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4001, Australia, k.mengersen@qut.edu.au*

[7] *Centre d'Ecologie Fonctionnelle et Evolutive - UMR 5175 - CNRS, EPHE-PSL University, Université Paul Valéry Montpellier 3, Montpellier, 34293, France, Aurelien.BESNARD@cefe.cnrs.fr*

Résumé. L'étude de la distribution des espèces le long de gradients environnementaux est très répandue en écologie et évolution. Il est également courant de s'intéresser aux changements de distribution entre deux périodes, notamment dans le contexte des changements globaux. Spécifiquement, l'étude de changement d'optimums de répartition a fait l'objet d'une large littérature. Jusqu'à présent les déplacements d'optimums ne sont pas estimés explicitement par les modèles utilisés. Nous proposons un nouveau modèle Bayésien permettant d'estimer explicitement et simultanément pour plusieurs espèces le déplacement d'optimums. Plusieurs scénarios de simulation ont permis d'évaluer son efficacité et de la comparer à deux méthodes souvent utilisées : le modèle linéaire généralisé mixte et la comparaison de moyenne. Une étude sur les déplacements altitudinaux de 24 espèces d'orthoptères dans les alpes françaises, dans un contexte de changement climatique, a aussi permis d'illustrer les différences entre méthodes.

Mots-clés. Modèle hiérarchique, Bayésien, Modèle de distribution d'espèces, Déplacements altitudinaux, Changements globaux, Niche écologique

Abstract. Study of species response curves along environmental gradient is of great interest in ecology and evolution. Particularly in the context of global change, it has became recurrent to test for changes in species distribution between two surveys. Facing

a lack of an explicit method to estimate shift in species optimum, we extended an existing hierarchical Bayesian model to explicitly estimate multi-species optimum shifts. Through a simulation study, including different ecological scenarios, we assessed the performance of our new model and compared it with methods commonly used: Generalized Linear Mixed Model and mean comparison method. We also conducted a case study on Orthoptera community using the three methods. We investigated the altitudinal shifts of 24 Orthoptera species between two surveys apart from thirty years in the French Alps.

Keywords. Hierarchical model, Bayesian, Species distribution model, Altitudinal shift, Global change, Ecological niche

1 Context

Modelling species distribution along environmental gradient is of particular interest in ecology and evolution. Particularly in the context of global change, it has become recurrent to test for changes in species distribution between two surveys. Studies investigating shifts in species distribution are numerous. A lot of them focused on changes in range limits, though it has been criticized as bias could easily occur due to sampling artefact (Shoo, Williams, and Hero 2006). Hence, several authors had preferred studying shifts in species optimum positions as its less dependent on sampling effort (Shoo, Williams, and Hero 2006).

Different modelling frameworks were developed for studying optimum shifts. The simplest and widely used, is the comparison of species' mean position along environmental gradients (Chen et al. 2009; Menéndez et al. 2014; Freeman et al. 2018). However, concern has been raised about the effect of sampling effort on such method (Shoo, Williams, and Hero 2006; Ter Braak and Looman 1986). Statistical models, such as generalized linear model or generalized additive model, are also commonly used (Lembrechts et al. 2017; Tayleur et al. n.d.). In most of the studies, two models were conducted separately to estimate species optimums for each period. Then the shift was derived. One gap of this type of analysis is the lack of uncertainty associated with the estimate. To face it some authors used bootstrapping (Maggini et al. 2011), though it has some inconveniences (Urli et al. 2014).

We propose here an extension of the Gaussian logistic model (Jamil, Kruk, and Braak 2014; Ter Braak and Looman 1986) that explicitly estimates shifts in optimums of multiple species. We also developed a Bayesian generalized linear mixed model and used the posterior distributions of its coefficients to compute posterior distributions of species optimum shifts. These hierarchical Bayesian models were evaluated and compared to the mean comparison method through a simulation study involving different sampling and ecological scenarios. We also present the results of the motivating example which

consisted of estimating species-specific altitudinal shifts of an Orthopteran community between two surveys apart from thirty years.

2 Optimum shift modelling

Our first purpose was to study the motivating example. We wanted to estimate species-specific shifts in optimum positions along an altitudinal gradient of Orthoptera species between two surveys. Hence, for clarity we stayed in the context of altitudinal shifts between two surveys. However, the methods developed could be applied in other contexts such as ontogenetic, geographical or phenological shifts to cite a few (Bertrand, Gégout, and Bontemps 2011; Coudun and Gégout 2005; Hällfors et al. 2020).

2.1 Mean comparison method

Positions of species optimums along the altitudinal gradient are seen, in the mean comparison method (coded hereafter T-test), as the averages of altitudes at which species were observed. We conducted two-sample Student tests between surveys on the occupied altitudes, for each species separately. Hence, it gave us species-specific shift estimates and confidence intervals associated.

2.2 Bayesian models

We developed two Bayesian hierarchical models: a Generalized Linear Mixed Model (GLMM) and a new model, named Explicit Hierarchical Model of Optimum Shifts (EHMOS). In the GLMM, the species-specific occupancy probabilities were modelled as a logistic regression of linear and quadratic effects of altitude, linear effect of the survey and their interactions, allowing for variability in species response curves between surveys:

$$logit(\psi_{i,j,s}) = \beta_{0,i} + \beta_{1,i} \times X_j + \beta_{2,i} \times {X_j}^2 + \beta_{3,i} \times S_s + \beta_{4,i} \times S_s \times X_j + \beta_{5,i} \times S_s \times {X_j}^2$$

where $\beta_i' s$ are the coefficients, coded as species random effects, related to the species-specific effects mentioned above and X_j the altitude at site j. S_s is the binary variable coding for the survey. Posterior distributions of species shifts were calculated from posterior distributions of the model coefficients. The EHMOS is an extension at two surveys of the multi-species Gaussian logistic model developed to describe species unimodal response curves for one survey (Jamil, Kruk, and Braak 2014). Our model explicitly estimate species shifts in optimum positions between two surveys. Species response curves were described through a reformulated logistic regression involving four ecologically meaningful parameters ($\alpha_{i,s}$, θ_i, δ_i and $\tau_{i,s}$):

$$logit(\psi_{i,j,s}) = \alpha_{i,s} - \frac{(\mathrm{X}_j - (\theta_i + (\mathrm{S}_s \times \delta_i)))^2}{2 \times {\tau_{i,s}}^2} \quad (1)$$

where $\alpha_{i,s}$ represents the maximum occupancy probability on the logit scale reach by species i during survey s; θ_i is the environmental optimum of species i for the first survey (as the variable S is coded as $S_1 = 0$ and $S_2 = 1$); δ_i is the shift between the two optimums for species i; $\tau_{i,s}$, named tolerance, is a measure of response curve width of species i in survey s. Each parameter is seen as a random effect varying between species. The maximum occupancy probability and the tolerance could also vary between surveys. In both model, we computed medians and credible intervals of the posterior distributions of estimated optimum shifts.

3 Motivating example

We estimated shifts in optimum positions along an altitudinal gradient for 24 Orthoptera species between one historic survey, conducted between 1983 and 1988 (Gueguen 1990), and one recent survey, conducted in 2018 and 2019 (Mourguiart et al. 2021). 134 sites distributed between 928 m and 2614 m of altitude were sampled in both surveys. The three statistical methods provided quite different estimates, but all suggested an upward shift in average despite heterogeneity between species.

4 Simulation study

Through a simulation study including eight scenarios, we tested the performance of the EHMOS and compared it to the GLMM and the mean comparison method. The eight scenarios were decomposed in three sub-scenarios referring to the (A) sampling design, (B) species optimum positions, and (C) species ecological specializations. Each sub-scenario was cut in two categories. For each category, unimodal symmetric response curve (Equation 1) of twenty species were simulated along a virtual altitudinal gradient. Presence/absence data were derived from the response curves for 300 simulated sampling sites. The sampling sites were either uniformly or normally distributed along the virtual gradient depending on the sampling sub-scenario category (A1 or A2 respectively). Once the sampling sites were simulated, we positioned species optimums on the sampling gradient according to the optimum sub-scenario category (coded B1 and B2). In sub-scenario B1, 100% of species have their both optimums in the middle of the sampling range, while in B2 30% of species had their optimums close to the sampling boundaries. For all scenarios, we computed 60% of upward shifts, 20% of downward shifts and 20% of no shift to test all kinds of response while keeping majority of upward shifts as expected under ongoing climate change. Finally, we simulated species response curve shape descriptors following two ecological specialization sub-scenario categories (C1 and C2). In sub-scenario C1, we simulated all species as specialist species, *i.e.* species having narrow ecological niche width and high maximum probability of occurrence. In sub-scenario C2, we simulated 50% of specialist and 50% of generalist species with wider niche width and smaller max-

imum probability of occurrence. We run all possible combinations of the sub-scenario categories resulting in eight scenarios (2^3 scenarios). Model performance was assessed computing bias and RMSE. Results indicate that EHMOS provides the best estimates for all scenarios, with the lowest RMSEs (Table 1). The mean comparison method is the poorest in almost all scenarios, mainly due to underestimation. GLMM seems to provide quite good estimates but gave unreasonable estimates for edge species especially under uniform sampling and had some convergence issues.

Table 1: Comparison of estimation accuracy, based on simulated data analyses, between our new model, EHMOS, explicitely estimating species shift, a simple GLMM and a mean comparison method. The table shows results obtained from data simulated according to eight scenarios. For each model and each scenario, we show the average root-mean-square error ('RMSE') and the bias.

	RMSE			Bias		
Scenario	EHMOS	GLMM	T-test	EHMOS	GLMM	T-test
A1xB1xC1	10.65	21.24	67.81	0.79	-0.74	-3.31
A1xB1xC2	29.12	34.60	79.19	2.06	-1.56	3.13
A1xB2xC1	42.89	273.59	131.31	0.66	8.00	1.93
A1xB2xC2	96.87	135.48	147.28	-5.03	11.46	-12.08
A2xB1xC1	8.85	11.14	75.52	0.73	0.62	-5.00
A2xB1xC2	22.97	36.71	193.70	0.38	1.74	-24.84
A2xB2xC1	15.66	41.17	80.25	1.81	-0.19	-8.40
A2xB2xC2	59.52	120.18	145.15	0.52	-1.70	-17.17

We show in this study that the new Explicit Hierarchical Model of Optimum Shifts (EHMOS) provide better estimates than GLMM and T-test methods commonly used in range shift studies. The explicit parametrization of optimum shifts in EHMOS could allow new insights in shift modelling. For instance, we could directly incorporate potential effects of species ecological traits on optimum shifts, which is currently studied a posteriori (Felde, Kapfer, and Grytnes 2012).

Bibliographie

Bertrand, R., Gégout, J-C., and Bontemps, J-D. (2011). Niches of temperate tree species converge towards nutrient-richer conditions over ontogeny. *Oikos*, 120.10, pp. 1479–1488.
Chen, I-C. et al. (2009). Elevation increases in moth assemblages over 42 years ona tropical mountain. *Proceedings of the National Academy of Sciences*, 106.5, pp. 1479–1483.
Coudun, C. and Gégout, J-C. (2005). Ecological behaviour of herbaceous forest species along a pH gradient: a comparison between oceanic and semicontinental regions in northern France. *Global Ecology and Biogeography*, 14.3, pp. 263–270.

Felde, V. A., Kapfer, J., Grytnes, J-A. (2012). Upward shift in elevational plant species ranges in Sikkilsdalen, central Norway. *Ecography*, 35.10, 922-932.
Freeman, B. G. et al. (2018). Climate change causes upslope shifts and mountaintop extirpations in a tropical bird community. *Proceedings of the National Academy of Sciences*, 115.47, pp. 11982–11987.
Gueguen, A. (1990). Impact du pâturage ovin sur la faune sauvage: exemple des Orthoptères.
Hällfors, M. H. et al. (2020). Shifts in timing and duration of breeding for 73 boreal bird species over four decades. *Proceedings of the National Academy of Sciences*, 117.31, pp. 18557–18565.
Jamil, T., Kruk, C., and ter Braak, C. JF. (2014). A Unimodal Species Response Model Relating Traits to Environment with Application to Phytoplankton Communities. *PLOS ONE*, 9.5, pp. 1–14.
Lembrechts, J J. et al. (2017). Mountain roads shift native and non-native plant species' ranges. *Ecography*, 40.3, pp. 353–364.
Maggini, R et al. (2011). Are Swiss birds tracking climate change?: Detecting elevational shifts using response curve shapes. *Ecological Modelling*, 222.1, pp. 21–32.
Menéndez, R. et al. (2014). Climate change and elevational range shifts: evidence from dung beetles in two European mountain ranges. *Global Ecology and Biogeography*, 23.6, pp. 646–657.
Mourguiart, B., et al. (2021), Multi-species occupancy models: an effective and flexible framework for studies of insect communities. *Ecological Entomology*, 46.2, pp 163-174.
Schleuning, M., et al. (2020), Trait-Based Assessments of Climate-Change Impacts on Interacting Species. *Trends in Ecology and Evolution*, 35.4, pp. 319-328.
Shoo, L. P., Williams S. E., and Hero, J-M. (2006). Detecting climate change induced range shifts: Where and how should we be looking? *Austral Ecology*, 31.1, pp. 22–29.
Tayleur, C. et al. (2015). Swedish birds are tracking temperature but not rainfall:evidence from a decade of abundance changes. *Global Ecology and Biogeography*, 24.7, pp. 859–872.
Ter Braak, C. JF and Looman C. WN (1986). Weighted averaging, logistic regression and the Gaussian response model. *Vegetation*, 65.1, pp. 3–11.
Urli, M. et al. (2014). Inferring shifts in tree species distribution using asymmetric distribution curves: a case study in the Iberian mountains. *Journal of Vegetation Science*, 25.1, pp. 147–159.

Link between Threshold ARMA and tdARMA models

Guy Mélard [1] & Marcella Niglio [2]

[1] *Université Libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Bruxelles, Belgique, gmelard@ulb.ac.be*

[2] *Universitá degli Studi di Salerno, Department of Economics and Statistics, Via Giovanni Paolo II, 132, Fisciano (SA), Italy, mniglio@unisa.it*

Résumé. Dans la présente contribution nous proposons un lien entre les modèles autorégressifs moyenne mobile à seuils (Threshold Autoregressive Moving Average) ou TARMA et les modèles ARMA dépendant du temps (Time-Dependent ARMA) ou *td*ARMA. Nous montrons qu'une paramétrisation adéquate permet d'inclure les modèles TARMA dans la large classe des structures *td*ARMA. Le principal advantage qu'on peut tirer de ce résultat est l'obtention des propriétés asymptotiques des estimateurs des paramètres TARMA sous des conditions plus faibles que celles disponibles dans la litérature.

Mots-clés. Modèle à seuils, modèle ARMA dépendant du temps.

Abstract. In the present contribution, we propose a link between Threshold Autoregressive Moving Average (TARMA) and Time-Dependent ARMA (*td*ARMA) models. We show that a proper parametrization allows to include the TARMA model in the large class of *td*ARMA structures. The main advantage that can be obtained from this result is the derivation of the asymptotic properties of the estimators of TARMA parameters that can be obtained under weaker conditions with respect to those in the available literature.

Keywords. Threshold model, time-dependent ARMA model.

1 Introduction

In time series analysis, the dynamic of data has often been modelled through the introduction of time-dependent coefficient structures. Starting from Nicholls and Quinn (1982) different proposals have been given in this domain.

In the present contribution, the attention is focused on some generalizations of the ARMA structure where the dependence of the coefficients to the time is differently modelled.

More precisely we will start considering Threshold ARMA (TARMA) models (Tong, 1983):

$$X_t = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \phi_j^{(i)} X_{t-j} - \sum_{j=1}^{q} \theta_j^{(i)} \epsilon_{t-j} \right] \mathbb{I}_{\{Y_{t-d} \in \mathbb{R}_i\}} + \epsilon_t, \tag{1}$$

where X_t is the variable of interest at time t, k is the number of regimes, the $\phi_j^{(i)}$, $j = 1, ..., p$, and $\theta_j^{(i)}$, $j = 1, ..., q$, are, respectively, the autoregressive and moving average coefficients of the ARMA models for the i-th regime, $i = 1, ..., k$, Y_{t-d} is the threshold variable, d is the threshold delay, $\mathbb{I}_{\{\cdot\}}$ is an indicator function, $\mathbb{R}_i$ is a subset of the real line such that $\mathbb{R} = \bigcup_{i=1}^{k} \mathbb{R}_i$ with $\mathbb{R}_i \cap \mathbb{R}_s = \emptyset$, for $i \neq s$, and $\{\epsilon_t\}$ a sequence of independent and identically distributed (i.i.d.) random variables with null mean and finite moments of order $4+\delta$, $\delta > 0$, with ϵ_t independent from Y_t and Y_t a stationary and ergodic process.

Even if model (1) can be shortly described as "local linear ARMA" because, within each regime, X_t follows an ARMA model, its overall structure is more complex and goes beyond the linear domain. This is the reason why general results for the statistical properties of model (1), such as stationarity and ergodicity, have only been faced for well-defined parametrizations with endogenous threshold variable ($Y_{t-d} = X_{t-d}$): Brockwell *et al.* (1992), consider a simplified structure with $\theta_j^{(i)} = \theta_j$, for $j = 1, \ldots, q$ where, in other words, the moving average coefficients do not change among regimes; Liu and Susko (1992) define sufficient conditions for the stationarity of model (1) with $p = 1$ whereas more recently Chan and Goracci (2019) focus the attention on the ergodicity of first-order threshold ARMA processes (with $p = q = 1$).

As clarified before, in all cited literature the examined threshold model is characterized by an endogenous threshold variable. It makes, at the same time, the model less general with respect to model (1) but even more complex, when its dynamic structure needs to be investigated.

A recent contribution in this domain is given in Boubacar Maïnassara and Rabehasaina (2020), where, differently from model (1) the switching structure at known dates is related to an observed process with values in a finite set.

In the following, we consider a further variant of (1) that allows connecting the TARMA model to the large class of *td*ARMA models (Azrak and Mélard, 1998). We are going to present the model, the main differences with respect to model (1), and how these differences can support the estimation of the model parameters.

2 Threshold ARMA model

In (1), t is assumed to vary in $\mathbf{Z}$, the set of integers. In the following we consider the process X_t starting at time $t = 1$, such that $X_t = 0$ and $\epsilon_t = 0$ for $t < 1$.

A *td*ARMA model is defined by

$$X_t = \sum_{j=1}^{p} \phi_{tj} X_{t-j} - \sum_{j=1}^{q} \theta_{tj} \epsilon_{t-j} + \epsilon_t, \tag{2}$$

where the coefficients ϕ_{tj}, $j = 1, \ldots, p$, and θ_{tj}, $j = 1, \ldots, q$, depend on a vector of parameters $\boldsymbol{\beta}$ and ϵ_t is like above.

To better understand the relation between the TARMA and the tdARMA models, we introduce the following notation: let $\mathbb{I}_{t-d}^{(i)}$ be a short form for the indicator function $\mathbb{I}_{\{Y_{t-d}\in\mathbb{R}_i\}}$, such that $\mathbb{I}_{t-d}^{(i)}=1$ if $y_{t-d}\in\mathbb{R}_i$ and $\mathbb{I}_{t-d}^{(i)}=0$ otherwise, for $i=1,\ldots,k$, model (1) can be written as (2) where

$$\phi_{tj}(\boldsymbol{\beta})=\sum_{i=1}^{k}\phi_j^{(i)}\mathbb{I}_{t-d}^{(i)},\quad j=1,\ldots,p,$$

$$\theta_{tj}(\boldsymbol{\beta})=\sum_{i=1}^{k}\theta_j^{(i)}\mathbb{I}_{t-d}^{(i)},\quad j=1,\ldots,q,$$

and $\boldsymbol{\beta}=(\phi_1^{(1)},\cdots,\phi_p^{(k)},\theta_1^{(1)},\cdots,\theta_q^{(k)})$, with $\boldsymbol{\beta}\in\mathbf{B}$ an open set of a Euclidean space $\mathbf{R}^{(p+q)k}$ and let $\boldsymbol{\beta}_0$ (an interior point of $\mathbf{B}$) be the corresponding vector of the true parameters. Let $e_t(\boldsymbol{\beta})$ be the residual defined iteratively by

$$X_t=\sum_{i=1}^{k}\left[\sum_{j=1}^{p}\phi_j^{(i)}X_{t-j}-\sum_{j=1}^{q}\theta_j^{(i)}e_{t-j}(\boldsymbol{\beta})\right]\mathbb{I}_{\{Y_{t-d}\in\mathbb{R}_i\}}+e_t(\boldsymbol{\beta}),\tag{3}$$

for $t=1,2,....$ Following Francq and Gautier (2004) and the notation of Mélard (2021), model (2) can be iteratively given as:

$$\mathbf{X}_t(\boldsymbol{\beta})=\sum_{r=0}^{t-1}\left[\sum_{s=0}^{r}\left(\prod_{\ell=0}^{r-s-1}\mathbf{J}\mathbf{A}_{t-\ell}(\boldsymbol{\beta})\right)\mathbf{K}\left(\prod_{j=0}^{s-1}\mathbf{A}_{t-r+s-j}(\boldsymbol{\beta}_0)\right)\right]\mathbf{E}_{t-r},\tag{4}$$

where:

$$\underset{[(p+q)\times1]}{\mathbf{X}_t}(\boldsymbol{\beta})=\begin{bmatrix}X_t\\ \vdots\\ X_{t-p+1}\\ e_t(\boldsymbol{\beta})\\ \vdots\\ e_{t-q+1}(\boldsymbol{\beta})\end{bmatrix},\quad \underset{[(p+q)\times(p+q)]}{\mathbf{A}_t(\boldsymbol{\beta})}=\begin{bmatrix}\boldsymbol{\Phi}_t & \tilde{\boldsymbol{\Theta}}_t\\ \underset{(q\times p)}{\mathbf{0}} & \mathbf{C}\end{bmatrix},\quad \underset{[(p+q)\times1]}{\mathbf{E}_t}=\begin{bmatrix}\epsilon_t\\ \underset{[(p-1)\times1]}{\mathbf{0}}\\ \epsilon_t\\ \underset{[(q-1)\times1]}{\mathbf{0}}\end{bmatrix},$$

with:

$$\boldsymbol{\Phi}_t=\begin{bmatrix}\phi_{t1} & \phi_{t2} & \cdots & \phi_{tp}\\ & \underset{(p-1)}{\mathbf{I}} & & \underset{(p-1)\times1}{\mathbf{0}}\end{bmatrix},\quad \tilde{\boldsymbol{\Theta}}_t=\begin{bmatrix}\theta_{t1} & \theta_{t2} & \cdots & \theta_{tq}\\ & \underset{(q-1)\times q}{\mathbf{0}} & &\end{bmatrix},\quad \mathbf{C}=\begin{bmatrix}\underset{[1\times(q-1)]}{\mathbf{0}} & 0\\ \underset{(q-1)}{\mathbf{I}} & \underset{[(q-1)\times1]}{\mathbf{0}}\end{bmatrix},$$

whereas $\mathbf{I}$ is the identity matrix, $\mathbf{0}$ a null vector or matrix, $\mathbf{K}=[k_{u,v}]$, for $u,v=1,\ldots,(p+q)$, is a null matrix with two elements replaced with $k_{1,1}=k_{(p+1),1}=1$ and $\mathbf{J}=[j_{u,v}]$ is an identity matrix with two elements replaced with $j_{1,1}=0$ and $j_{(p+1),1}=-1$.

Given these results, model (4) can be shortly written as:

$$\mathbf{X}_t(\boldsymbol{\beta}) = \sum_{r=0}^{t-1} \Psi_{tr}(\boldsymbol{\beta})\mathbf{E}_{t-r}, \tag{5}$$

with $\Psi_{tr}(\boldsymbol{\beta}) = \sum_{s=0}^{r} \left(\prod_{\ell=0}^{r-s-1} \mathbf{J}\mathbf{A}_{t-\ell}(\boldsymbol{\beta})\right) \mathbf{K} \left(\prod_{j=0}^{s-1} \mathbf{A}_{t-r+s-j}(\boldsymbol{\beta}_0)\right)$.

Finally note that the results given in this section for the TARMA model with exogenous threshold variable can be applied to the case where the threshold variable is endogenous, $Y_{t-d} = X_{t-d}$, and, in our knowledge, it is a novelty that has not been considered in the cited literature. Note also that this section (and the next one) can be written also for a vector TARMA model (VTARMA), see e.g. Niglio and Vitale (2015).

3 MA representation of the TARMA model

The notation introduced in Section 2 allows obtaining the MA representation of $e_t(\boldsymbol{\beta})$. In fact, noting that $e_t(\boldsymbol{\beta})$ is the $(p+1)$-th element in $\mathbf{X}_t(\boldsymbol{\beta})$, then:

$$e_t(\boldsymbol{\beta}) = \sum_{r=1}^{t-1} \psi_{tr}(\boldsymbol{\beta})\epsilon_{t-r}. \tag{6}$$

with $\psi_{tr}(\boldsymbol{\beta}) = \mathbf{U}'_{p+1}\Psi_{tr}(\boldsymbol{\beta})\mathbf{U}_1$, where $\mathbf{U}_1$ and $\mathbf{U}_{p+1}$ are two $[(p+q)\times 1]$ null vectors with the first and the $(p+1)$-th elements replaced with 1 respectively and $\mathbf{U}'_{p+1}$ the transpose of $\mathbf{U}_{p+1}$.

It is then possible to obtain the first three derivatives of $e_t(\boldsymbol{\beta})$ with respect to the elements of $\boldsymbol{\beta}$, e.g. for the first-order derivative with respect to β_i, $i = 1, ..., m$, where $m = (p+q)k$,

$$\frac{\partial e_t(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{r=1}^{t-1} \psi_{tir}(\boldsymbol{\beta})\epsilon_{t-r}, \tag{7}$$

where

$$\psi_{tir}(\boldsymbol{\beta}) = \frac{\partial \psi_{tr}(\boldsymbol{\beta})}{\partial \beta_i}. \tag{8}$$

Then we let $\psi_{tir} = \psi_{tir}(\boldsymbol{\beta}_0)$. Starting from here, we restrict the TARMA model to the case of a strictly exogenous variable Y_{t-d}. Otherwise, in the case of a TARMA with an endogenous threshold variable, the coefficients ψ_{tir} will be random variables.

Given the first three derivatives of $e_t(\boldsymbol{\beta})$ like in (7) and using the results in Francq and Gautier (2004) and Mélard (2021) for homoscedastic *td*VARMA models, we can obtain, under the conditions A1-A6 in Mélard(2021), see the Appendix, quasi maximum likelihood estimators for $\boldsymbol{\beta}$ whose asymptotic properties are derived.

These conditions can be checked under relatively weak assumptions on the TARMA model except A4. Indeed, A1 and A3 are trivially true, and it can be seen that A2, A5,

and A6 are verified if the k ARMA models involved in (1) are stationary and invertible, by proceeding like in Mélard (2021). This is because the Frobenius norm of the products in (4) can then be bounded by a power of some constant $\Phi < 1$ by using properties of the eigenvalues of companion matrices.

It is unfortunately not possible to check (A4): it remains a condition. A natural requirement is that the k ARMA models involved in (1) are identifiable (so each of them has no common root for their autoregressive and moving average polynomials) but it is not enough to guarantee the existence and invertibility of the information matrix.

References

Azrak R. and Mélard G. (1998), The exact quasi-likelihood of time-dependent ARMA models *Journal of Statistical Planning and Inference*, 68, pp. 31-45.

Boubacar Maïnassara Y. and Rabehasaina L. (2020), Estimation of weak ARMA models with regime changes *Statistical Inference for Stochastic Processes*, 23, pp. 1-52.

Brockwell P.J., Liu J. and Tweedie R.L. (1992), On the existence of stationary threshold autoregressive moving-average processes, *Journal of Time Series Analysis*, 13, pp. 95-107.

Chan K.S. and Goracci G. (2019), On the ergodicity of first-order threshold autoregressive moving-average processes, *Journal of Time Series Analysis*, 40, pp. 256-264.

Francq C. and Gautier A. (2004), Estimation of time-varying ARMA models with Markovian changes in regime *Statistics & Probability Letters*, 70, pp. 243-25.

Liu J. and Susko E. (1992), On strict stationarity and ergodicity of a non-linear ARMA model, *Journal of Applied Probability*, 29, pp. 363-373.

Mélard G. (2021), An indirect proof for the asymptotic properties of VARMA model estimators, *Econometrics and Statistics*, in press. https://doi-org/10.1016/j.ecosta.2020.12.004

Nicholls, D.F. and Quinn, B.G. (1982), *Random Coefficient Autoregressive Models: An Introduction*, Springer-Verlag, New York.

Niglio, M. and Vitale, C.D. (2015), Threshold vector ARMA models, *Communications in Statistics: Theory and Methods*, 44(14), pp. 2911-2923.

Tong, H. (1983), *Threshold Models in Non-linear Time Series Analysis*, Springer-Verlag, New York.

Appendix: the assumptions of Mélard (2021)

.

We use the notations introduced in Sections 2 and 3. We consider a homoscedastic *td*VARMA model of dimension ℓ (equal to $p+q$ above) and denote Σ the covariance matrix of the innovations E_t supposed to be invertible (this is not the case here).

We suppose (A1) that the coefficient matrices $\Phi_{ti}(\beta)$ et $\Theta_{tj}(\beta)$ are of class C^3 in β in an open set B which contains the true value β_0; (A2) denoting $\|.\|_F$ the Frobenius norm of a matrix, that there exist positive constants N_l, $l = 1, ..., 4$, and $0 < \Phi < 1$ such that, for $\nu = 1, ..., t-1$, and $i = 1, ..., m$, $\sum_{r=1}^{t-1} \|\Psi_{tr}\|_F^2 < N_1$, $\sum_{r=1}^{t-1} \|\Psi_{tr}\|_F^4 < N_2$, $\sum_{r=\nu}^{t-1} \|\Psi_{tir}\|_F^2 < N_3\Phi^{\nu-1}$, $\sum_{r=\nu}^{t-1} \|\Psi_{tik}\|_F^4 < N_4\Phi^{\nu-1}$ and three others for second and third order derivatives; (A3) $\kappa = E\left(\mathrm{vec}(E_tE_t^T)\,\mathrm{vec}(E_tE_t^T)^T\right) = E\left((E_tE_t^T)\otimes(E_tE_t^T)\right)$ exists and does not depend on t; (A4) the limits $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E_{\beta_0}\left(\frac{\partial E_t^T(\beta)}{\partial\beta_i}\Sigma^{-1}\frac{\partial E_t(\beta)}{\partial\beta_j}\right) = V_{ij}$ exist for $i, j = 1, ..., m$, where the matrix $V = (V_{ij})_{1\le i,j\le m}$ is strictly positive definite; (A5) $\frac{1}{n^2}\sum_{d=1}^{n-1}\sum_{t=1}^{n-d}\sum_{r=1}^{t-1} \|\Psi_{tir}\|_F \left\|\Psi_{t+d,j,r+d}\right\|_F = O\left(\frac{1}{n}\right)$, $\quad i, j = 1, ..., m$; (A6) Similarly

$$\frac{1}{n^2}\sum_{d=1}^{n-1}\sum_{t=1}^{n-d}\left[\sum_{r=1}^{t-1} M_{t0rr}^{jiT}\Xi(\Sigma)M_{tdrr}^{ij} \quad + \quad \sum_{r_1=1}^{t-1}\sum_{r_2=1}^{t-1} M_{t0r_2r_1}^{jiT}K_{\ell,\ell}(\Sigma\otimes\Sigma)M_{tdr_1r_2}^{ij} \right.$$
$$\left. + \sum_{r_1=1}^{t-1}\sum_{r_2=1}^{t-1} M_{t0r_2r_1}^{jiT}(\Sigma\otimes\Sigma)M_{tdr_2k_1}^{ij}\right] = O\left(\frac{1}{n}\right),$$

with a commutation matrix $K_{\ell,\ell}$, $\Xi(\Sigma) = \kappa - \mathrm{vec}(\Sigma).\,\mathrm{vec}(\Sigma)^T - (\Sigma\otimes\Sigma) - K_{\ell,\ell}(\Sigma\otimes\Sigma)$, and, for $r', r'' = r, r_1, r_2$, $M_{tfr'r''}^{ij} = \mathrm{vec}(\Psi_{t+f,i,r'+f}^T\Sigma^{-1}\Psi_{t+f,j,r''+f})$, $f = 0, d$, $i, j = 1, ..., m$.

Estimation of copulas and their densities by projection with application in insurance

Yves I. Ngounou Bakam[1] & Denys Pommeret[1,2]

[1] *CNRS, Ecole Centrale, I2M, Aix-Marseille Univ, Marseille, France*
yves-ismael.ngounou-bakam@univ-amu.fr
[2] *ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, France*
denys.pommeret@univ-amu.fr

Résumé. Nous proposons un nouvel estimateur de la copule et de la densité de copule par des approximations de leurs projections orthogonales. Nous montrons que les marges de l'estimateur de la densité de copule sont uniformes et nous étudions les propriétés asymptotiques, notamment l'optimalité du point de vue minimax et maxiset. Une méthode de sélection automatique permet de choisir l'ordre d'approximation. Nous confrontons les performances de ces estimateurs aux performances des estimateurs existants dans la littérature. La méthode est illustrée sur un cas concret en assurance.

Mots-clés. Copule, Estimation non paramétrique, Minimax, Maxiset, Polynômes de Legendre.

Abstract. Non parametric estimators of copulas and copula densities are obtained by orthonornal projections. Their asymptotic properties are reviewed. We investigate the optimality of both procedure in the minimax sense and maxiset approach. We provide a data driven method to select the number of projections which allows these estimators to be well adapted to all types of copulas. In particular, this selection makes it possible to detect the cases of independence. A real dataset in insurance is studied.

Keywords. Copula estimators, Maxiset theory, Minimax theory, Legendre polynomials, Nonparametric estimation.

1. Introduction

Consider a d-random vector $\mathbf{X} = (X_1, \cdots, X_d)^T$ with joint distribution function H and marginal distribution functions $F_1, \cdots, F_d$, that we assume to be continuous. According to Sklar's Theorem ([5]), there exists a unique d-variate function C such that

$$H(x_1, \cdots, x_d) \;=\; C(F_1(x_1), \cdots, F_d(x_d)). \tag{1}$$

The function C is called the copula associated to $\mathbf{X}$. The copula is a joint distribution function on $I^d = [0,1]^d$, with uniform margins and satisfying $C(u_1, \cdots, u_d) = H(F_1^{-1}(u_1), \cdots, F_d^{-1}(u_d))$, where, for $j = 1, \cdots, d$, $F_j^{-1}(u_j) = \inf\{x_j; F_j(x_j) \geq u_j\}$, is

the quantile function of F_j. Assuming that for $j = 1, \cdots, d$, F_j is differentiable, we can express the joint density h of $\mathbf{X}$ as

$$h(x_1, \cdots, x_d) = c\left(F_1(x_1), \cdots, F_d(x_d)\right) \prod_{j=1}^{d} f_j(x_j), \tag{2}$$

where for $j = 1, \cdots, d$, f_j is the marginal density of X_j and where

$$c \;=\; \frac{\partial^d C}{\partial x_1 \cdots \partial x_d}, \tag{3}$$

is called the density copula of $\mathbf{X}$.

In this work we first propose to revisit the nonparametric method for estimating copula densities by considering an orthogonal shifted Legendre polynomials expansion. In this sense, our approach lies between the Legendre multiwavelet procedure [2] and the Bernstein method. More precisely, the basic idea developed here is to consider the transformations $U_j = F_j(X_j)$, for $j = 1, \cdots, d$, which yield to a vector of uniform random variables denoted by

$$\mathbf{U} = (U_1, \cdots, U_d)^T.$$

Its joint distribution function has the form

$$H_{\mathbf{U}}(u_1, \cdots, u_d) = H(F_1^{-1}(u_1), \cdots, F_d^{-1}(u_d)),$$

and clearly, $\mathbf{U}$ and $\mathbf{X}$ have the same structure of dependence with the same copula. We deduce that

$$h_{\mathbf{U}}(u_1, \cdots, u_d) = c(u_1, \cdots, u_d), \tag{4}$$

where $h_{\mathbf{U}}$ denotes the joint density of $\mathbf{U}$. This basic equality is the start of the construction of the density copula estimator, expressing c in a basis of orthogonal polynomials. In addition, a new copula estimator is derived by simply integrating the polynomial expansion. The properties of these estimators are studied. Finally, both estimators seem to outperform all others known in the statistical literature for a wide spectrum of scenarios. Moreover, these estimators are very simple and easy to implement and their execution time is very fast. Their interest is also demonstrated on a real dataset.

2. Construction of the estimators

Let us denote by μ the uniform measure on $I = [0, 1]$ and by $\{Q_m; m \in \mathbb{N}\}$ an orthonormal basis of shifted Legendre polynomials satisfying

$$\int Q_m(x) Q_k(x) \mu(dx) \;=\; \delta_{mk},$$

with $\delta_{mk} = 1$ if $m = k$ and 0 otherwise. For all $\mathbf{x} = (x_1, \cdots, x_d)^T \in I^d$ and for all $\mathbf{m} = (m_1, \cdots, m_d)^T \in \mathbb{N}^d$, we define

$$\mathbf{Q_m}(\mathbf{x}) = \prod_{j=1}^{d} Q_{m_j}(x_j),$$

and we write

$$\rho_{\mathbf{m}} = \mathbb{E}\left(\prod_{j=1}^{d} Q_{m_j}(F_j(X_j))\right). \tag{5}$$

Proposition 1. *Assume that*

$$\sum_{\mathbf{m}\in\mathbb{N}^d} \rho_{\mathbf{m}}^2 < \infty. \tag{6}$$

Then we have

$$c(u_1, \cdots, u_d) = \sum_{\mathbf{m}\in\mathbb{N}^d} \rho_{\mathbf{m}} \mathbf{Q_m}(\mathbf{u}) \tag{7}$$

$$C(u_1, \cdots, u_d) = \sum_{\mathbf{m}\in\mathbb{N}^d} \rho_{\mathbf{m}} \prod_{j=1}^{d} \int_0^{u_j} Q_{m_j}(x_j)\mu(dx_j). \tag{8}$$

It is important to note that the sequence $(\rho_{\mathbf{m}})_{\mathbf{m}\in\mathbb{N}^d}$ characterizes the copula. In this way it will be referred to as the *copula coefficients*. Since $Q_0 = 1$ we have $\rho_{\mathbf{0}} = 1$. The particular case $\rho_{\mathbf{m}} = 0$ for all $\mathbf{m} \neq 0$ coincides with the independent case. As seen in (5), the sequence $(\rho_{\mathbf{m}})_{\mathbf{m}\in\mathbb{N}^d}$ contains all the polynomial correlations between the marginal uniform random variables. In the bivariate case, that is when $d = 2$, the element $\rho_{\mathbf{m}}$ simply expresses the correlation between $Q_{m_1}(F_1(X_1))$ and $Q_{m_2}(F_2(X_2))$. In the general case, by orthogonality, we have $\mathbb{E}\left(Q_{m_j}(F_j(X_j))\right) = 0$ for all $m_j > 0$ and then $\rho_{\mathbf{m}} = 0$ as soon as $d-1$ components of $\mathbf{m}$ are equal to zero. Moreover, if for some integer $j \in \{1, \cdots, d\}$, X_j is independent to all over variables X_i, $i \neq j$, then $\rho_{\mathbf{m}} = 0$ as soon as $m_j > 0$. Then the copula coefficients can be used as an indicator of independence between the components of $\mathbf{X}$. In this sense, we also exhibit a link with the Spearman's rho.

For any positive integer vector $\mathbf{N} = (N_1, \cdots, N_d)^T$ we define the following $\mathbf{N}$-th order approximations:

$$c^{[\mathbf{N}]}(\mathbf{u}) = \sum_{\mathbf{m}\leq\mathbf{N}} \rho_{\mathbf{m}} \prod_{j=1}^{d} Q_{m_j}(u_j)$$

$$C^{[\mathbf{N}]}(\mathbf{u}) = \sum_{\mathbf{m}\leq\mathbf{N}} \rho_{\mathbf{m}} \prod_{j=1}^{d} \int_0^{u_j} Q_{m_j}(x_j)\mu(dx_j),$$

where the inequality $\mathbf{m} \leq \mathbf{N}$ means that $m_j \leq N_j$ for all $j = 1, \cdots, d$. We write $\mathbf{m} \not\leq \mathbf{N}$ when $\mathbf{m}$ does not satisfy this inequality. If we observe a n-sample $\mathbf{X}_1, \cdots, \mathbf{X}_n$, of iid random data, with $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})^T$ having joint distribution function H, then we can estimate the quantity $\rho_{\mathbf{m}}$ by

$$\widehat{\rho}_{\mathbf{m}} = \begin{cases} 1 & \text{if } \mathbf{m} = \mathbf{0}, \\ 0 & \text{if exactly } d-1 \text{ components of } \mathbf{m} \textit{ are zero}, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\prod_{j=1}^{d} Q_{m_j}(\widehat{F}_j(X_{ij})), & \text{else}, \end{cases} \tag{9}$$

where $\widehat{F}_j(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{1}(X_{ij} \leq x)$.

A $\mathbf{N}$-th order nonparametric estimator of the copula density c is given by

$$\widehat{c}^{[\mathbf{N}]}(u_1, \cdots, u_d) = \sum_{\mathbf{m} \leq \mathbf{N}} \widehat{\rho}_{\mathbf{m}} \prod_{j=1}^{d} Q_{m_j}(u_j). \tag{10}$$

By integration, we get a $\mathbf{N}$-th order nonparametric estimator of the copula function as follows

$$\widehat{C}^{[\mathbf{N}]}(u_1, \cdots, u_d) = \sum_{\mathbf{m} \leq \mathbf{N}} \widehat{\rho}_{\mathbf{m}} \prod_{j=1}^{d} \int_0^{u_j} Q_{m_j}(x_j)\mu(dx_j). \tag{11}$$

Note that the integration of Legendre polynomials and the uniform measure for μ yields to a very simple expression of (11).

3. Data driven bandwidth selection

We propose to use a data-driven procedure based on the Least-Squares Cross-Validation (LSCV) to select the optimal parameter $\widehat{\mathbf{N}}_{opt}$ ([6],[4]). The bandwidth parameter is the minimizer of the following function ([6],[4])

$$\widehat{LSCV}(\mathbf{N}) = \int_{I^d} \left(\widehat{c}^{[\mathbf{N}]}(\mathbf{u})\right)^2 d\mathbf{u} - \frac{2}{m}\sum_{i=1}^{m} \widehat{c}_{-i}^{[\mathbf{N}]}\left(\widehat{F}_1(X_{i1}), \cdots, \widehat{F}_d(X_{id})\right), \tag{12}$$

yielding to the following estimator of $\mathbf{N}$:

$$\widehat{\mathbf{N}}_{opt} = \underset{\mathbf{N} \in \mathbb{N}^d}{\operatorname{argmin}}\, \widehat{LSCV}(\mathbf{N}).$$

where $\widehat{c}_{-i}^{[\mathbf{N}]}$ is the leave-one-out copula density estimator without the data point $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})$.

The proposition below gives an abridged form of $\widehat{LSCV}(\mathbf{N})$ which is very useful to diminish its computation time in the numerical study.

Proposition 2.

$$\widehat{LSCV}(\mathbf{N}) = \frac{1}{n^2} \sum_{\mathbf{m} \leq \mathbf{N}} \left(\sum_{i=1}^{n} \prod_{j=1}^{d} Q_{m_j}^2 \left(\widehat{F}_j(X_{ij}) \right) - \frac{n+1}{n-1} \sum_{k \neq i} \prod_{j=1}^{d} Q_{m_j} \left(\widehat{F}_j(X_{ij}) \right) Q_{m_j} \left(\widehat{F}_j(X_{kj}) \right) \right).$$

4. Elementary properties of the estimators

Proposition 3. *Let $\mathbf{u} \in I^d$. The copula estimator $\widehat{C}^{[\mathbf{N}]}$ given by (11) satisfies the following properties*

i) $\widehat{C}^{[\mathbf{N}]}(\mathbf{u}) = 0$, *for all* $\mathbf{u} \in I^d$ *where at least one coordinate of* $\mathbf{u}$ *is zero.*

ii) If $\mathbf{u} = (1, \cdots, 1, u_i, 1, \cdots, 1)$ *then* $\widehat{C}^{[\mathbf{N}]}(\mathbf{u}) = u_i$.

iii) $\widehat{C}^{[\mathbf{0}]}(\mathbf{u}) = \prod_{j=1}^{d} u_j$.

Proposition 4. *The copula density estimator $\widehat{c}^{[\mathbf{N}]}$ given by (10) satisfies the following properties*

i) $\int_{I^d} \widehat{c}^{[\mathbf{N}]}(\mathbf{u}) \mu(d\mathbf{u}) = 1$ *and* $\int_{I^{d-1}} \widehat{c}^{[\mathbf{N}]}(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}_{-j}) = 1$ *with* $\mathbf{x}_{-j} = (x_1, \cdots, x_{j-1}, u_j, x_{j+1}, \cdots, x_d)$

iii) $\widehat{c}^{[\mathbf{0}]}(\mathbf{u}) = 1$, *for all* $\mathbf{u} \in I^d$.

Theorem 1. *Let (X, Y) be a continuous bivariate random variable with copula C. Then the Spearman's rho, namely ρ_C coincides with the first copula coefficient defined by (5), that is:*

$$\rho_C = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 = \rho_{(1,1)}$$

We can immediately deduce an estimator of the Spearman's rho as follows:

$$\hat{\rho}_C = \hat{\rho}_{(1,1)} = \frac{3}{n} \sum_{i=1}^{n} (2R_i - 1)(2S_i - 1).$$

where R_i and S_i are the rank statistics of X_i and Y_i respectively. This estimator is new and could be compared to the ones given in [3].

5. Minimax and maxiset results

Detailed Optimality results in the minimax/maxiset sense are available in [1].

Theorem 2. *Let* $v_n = n^{-\frac{b}{b+2d}}$, $\kappa > 0$ *and* $(\beta_1, \cdots, \beta_d) \in \mathbb{R}_+^d$. *Then the maxiset of* $\widehat{c}$ *for the rate of convergence* v_n, *denoted by* $\mathcal{MS}(\widehat{c}, v_n)$ *is given by*

$$\mathcal{MS}(\widehat{c}, v_n) = \left\{ c \in \mathbb{L}^2([0,1]^d) \ / \ \sum_{\mathbf{m} \in \mathbb{N}^d} \rho_{\mathbf{m}}^2 \left(1 + \sum_{j=1}^{d} m_j^{\beta_j} \right) < \kappa v_n \right\}$$

An intensive simulation study shows the very good performance of both copulas and copula densities estimators in comparison to a large panel of competitors. Main results, numerical experiments and real data application can be found in [1].

6. Extension

These estimators will allow up new aspects of copula statistics, namely the construction of equality tests. More precisely, we want to test hypothesis

$$H_0 : C^{\mathbf{X}} = C^{\mathbf{Y}} \text{ against } H_1 : C^{\mathbf{X}} \neq C^{\mathbf{Y}}.$$

where $C^{\mathbf{X}}$ and $C^{\mathbf{Y}}$ are copulas associated to $\mathbf{X} = (X_1, \cdots, X_d)$ and $\mathbf{Y} = (Y_1, \cdots, Y_d)$ respectively. By the previous expansions, it is clear that H_0 is equivalent to

$$H_0 : \rho_j^{\mathbf{X}} = \rho_j^{\mathbf{Y}}, \ \forall j \in \mathbb{N}^d \backslash \{\mathbf{0}_d\}$$

We will test this equality using (9) and the problem can be seen as a multivariate sample problem but with a copula point of wiew.

References

[1] Bakam, Y. I. N., and Pommeret, D. Nonparametric estimation of copulas and copula densities by orthogonal projections. *arXiv preprint arXiv:2010.15351* (2020).

[2] Chatrabgoun, O., Parham, G., and Chinipardaz, R. A legendre multiwavelets approach to copula density estimation. *Statistical Papers 58*, 3 (2017), 673–690.

[3] Pérez, A., and Prieto-Alaiz, M. A note on nonparametric estimation of copula-based multivariate extensions of spearman's rho. *Statistics & Probability Letters 112* (2016), 41–50.

[4] Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* (1982), 65–78.

[5] Sklar, A. Fonction de répartition dont les marges sont données. *Inst. Stat. Univ. Paris 8* (1959), 229–231.

[6] Taylor, C. C. Orthogonal series estimators and cross-validation. *Journal of Statistical Computation and Simulation 37*, 3-4 (1990), 151–158.

Identifying Probabilistic Anomalies Using Bayesian Networks In Presence Of Determinism

Raphaël Nedellec[1] & Tarek Benkhelif[2] & Evan Dufraisse[3] & Philippe Leray[4]

[1] *Talend, 89 Boulevard de la Prairie au Duc, 44200 Nantes, France, rnedellec@talend.com*
[2] *Talend, 89 Boulevard de la Prairie au Duc, 44200 Nantes, France, tbenkhelif@talend.com*
[3] *DUKe, LS2N, Université de Nantes, France, evan.dufraisse@imt-atlantique.net*
[4] *DUKe, LS2N, Université de Nantes, France, philippe.leray@ls2n.fr*

Résumé. Talend développe des solutions de traitements de données massives et d'ingestion de données pour ses utilisateurs. Il est important pour eux de pouvoir rapidement s'assurer de la qualité des données manipulées, de détecter des données abérantes, et de mettre en place des traitements adaptés. Différents produits (Talend Data Preparation, Talend Data Stewardship, Talend Pipeline Designer) permettent aux utilisateurs de mettre en place manuellement des procédure de nettoyage et de préparation de données. Notre objectif est d'accélerer la détection d'anomalies en découvrant de manière automatique des potentielles erreurs probabilistes. Nous développons une procédure de détection d'anomalies basée sur l'apprentissage d'un réseau bayésien en prenant en compte la présence potentielle de relations déterministes dans les jeux de données analysés.

Mots-clés. Réseaux bayésiens, déterminisme, détection d'anomalies, apprentissage statistique

Abstract. Talend develops software solutions to manipulate and ingest massive amount of data for its users. It is very important for them to quickly ensure the data quality and to detect anomalies and deal with those accordingly. Several products (Talend Data Preparation, Talend Data Stewardship, Talend Pipeline Designer) allows the users to manualy build pipelines to clean and transform data. Our main goal is to accelerate anomaly detection by automatically discovering probabilistic anomalies. We develop a procedure of anomaly detection based on learning a bayesian network, taking into account the possible presence of deterministic relationships within analysed datasets.

Keywords. Bayesian networks, determinism, anomaly detection, statistical learning

1 Data preparation and anomaly detection

Today, in every datascience project, a great amount of time is spent cleaning the data, and in particular removing and cleaning anomalies. The definition of what is an anomaly is in itself a challenge as an anomaly is often contextual. This is a reason why there

is a very vast litterature about anomaly detection, coming from both statistics and CS communities, as can be read in the following reviews [1,9,14]. For deterministic constraints violation, a great amount of work has been done to detect functional dependencies (FDs) errors and their extensions to non-canonical FDs [5]. For probabilistic anomalies, two approaches dominate: most of existing successful methods are either distance-based [2,4], or model-based [11,12]. In most of our customers' data, one can found an arbitrary mix of random and deterministic variables. We can face timeseries data, discrete or quantitative data, and even non structured textual data. Not all data are random, and a lot of data today is increasingly collected and generated by machines or software systems which rely very often on relational data models. This causes deterministic relationships between variables to be very common in our customers' datasets. One way to detect probabilistic anomalies is to learn a bayesian network model [7, 10]. Although we are interested in detecting both deterministic and probabilistic errors, our intent is to build a graphical model to detect probabilistic outliers given a prior knowledge of functional dependencies in our datasets.

2 Learning a bayesian network in presence of determinism to detect probabilistic anomalies

We decide to use bayesian networks which are a very flexible class of models that can represent very compactly the joint distribution of a set of variables. These models have been used to identify anomalies succesfully in various domains [3, 7, 10]. One challenge is to learn those models in presence of determinism between variables. Determinism can be seen as a degenerated relationship between tuple of variables. If $\mathbf{X}$ is a tuple of random variables, Y a simple random variable, and P a probability distribution over $(\mathbf{X}, Y)$, the relationship $\mathbf{X} \rightarrow Y$ is said to be deterministic iff there exists a function $f : Val(\mathbf{X}) \rightarrow Val(Y)$ such that $\forall(\mathbf{x}, y) \in Val(\mathbf{X}) \times Val(Y)$, $P(Y = y|\mathbf{X} = \mathbf{x}) = \mathbb{I}_{y=f(\mathbf{x})}$, i.e. iff for any realization $(\mathbf{x}, y)$ of $(\mathbf{X}, Y)$, we have $y = f(\mathbf{x})$ [13]. It has been shown that presence of determinism may induce false independence relations between variables and therefore inducing non faithful models. We compare several bayesian networks learning approaches to learn faithful models in presence of those deterministic relationships [8,13]. Another challenge remains in learning a robust bayesian network from real data, without having a clean dataset as input [6]. We will therefore present our procedure to detect probabilistic anomalies and show results for several datasets.

References

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. Detecting

data errors: where are we and what needs to be done? *Proceedings of the VLDB Endowment*, 9(12):993–1004, August 2016.

[2] Charu C Aggarwal and Philip S Yu. Outlier Detection for High Dimensional Data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 37–46, 2001.

[3] Batiste Le Bars and Argyris Kalogeratos. A Probabilistic Framework to Node-level Anomaly Detection in Communication Networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2188–2196, 2019. arXiv: 1902.04521.

[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jrg Sander. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[5] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. Relaxed Functional DependenciesA Survey of Approaches. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):147–165, January 2016.

[6] Yu Cheng, Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Robust Learning of Fixed-Structure Bayesian Networks. In *Advances in Neural Information Processing Systems*, pages 10283–10295, 2018.

[7] Kaustav Das and Jeff Schneider. Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, page 220, San Jose, California, USA, 2007. ACM Press.

[8] Sergio Rodrigues de Morais, Alexandre Aussem, and Marilys Corbex. Handling almost-deterministic relationships in constraint-based Bayesian network discovery: Application to cancer risk factor identification. In *16th European Symposium on Artificial Neural Networks*, Bruges, Belgium, April 2008.

[9] Victoria Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, October 2004.

[10] Romain Laby, Franois Roueff, and Alexandre Gramfort. Anomaly Detection and Localisation using Mixed Graphical Models. *arXiv:1607.05974 [stat]*, July 2016. arXiv: 1607.05974.

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, Pisa, Italy, December 2008. IEEE.

[12] Clement Pit-Claudel, Zelda Mariet, Rachael Harding, and Sam Madden. Outlier Detection in Heterogeneous Datasets using Automatic Tuple Expansion. Technical report, 2016.

[13] Thibaud Rahier. *Bayesian networks for static and temporal data fusion*. PhD thesis, Université Grenoble Alpes, 2018.

[14] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, October 2012.

Disentangling stellar-activity and planetary signals using Bayesian high-dimensional analysis.

Bo Ning [1]

[1] *Sorbonne Université, LPSM UMR 8001, 4, Place Jussieu 75252 Paris Cedex 05, France*

Résumé. Stellar activity, such as spots and faculae, provides a noise background that may lead to false discoveries or poor mass estimates of small planets when using radial velocity (RV) techniques, making it harder to detect Earth-like exoplanets. Spectroscopic activity indices are often used to verify the authenticity of planet candidates. In this talk, a Bayesian variable selection method (Ning et al., 2019) will be introduced that can automatically search for activity-sensitive lines through pixels from a set of spectra. Our method differs from those previously proposed methods, as it does not require the manual selection of a set of lines before conducting an analysis and can dependencies between lines are incorporated to the analysis. In the analysis, we not only consider the S-index—the most widely used activity indicator—but also include additional indices, including the Hα and NaD indices, the bisector inverse slope, and the full width at half maximum. Machine-readable tables and the code of the statistical method are available online. With stellar activity being the largest source of variability for next-generation RV spectrographs, this work could be a step toward accessing the myriad information available in high-precision spectra.

Mots-clés. Astrostatisics, Bayesian high-dimensional analysis, detecting exoplanets, variable selection

1 Introduction

Recent improvements in RV spectrograph technology are trying to bring instrumental noise levels below typical stellar noise levels in next-generation RV spectrographs, e.g., the EXtreme PREcision Spectrometer (EXPRES; Jurgenson et al. 2016). This leaves stellar activity as the largest source of variability. Stellar activity, including spots and faculae, can create velocity signals that confound planetary RV data.

In Ning et al. (2019), we propose a Bayesian variable selection method to identify activity-sensitive lines, using the sparse linear regression model. The covariates in this model are the pixels of the spectra and the response variable is an activity index. Unlike the regular linear regression model, a sparsity assumption is imposed on the regression coefficients. This assumption assumes that most regression coefficients will be very close

to zero or exactly zero (not activity-sensitive) and only a small number are non-zero (activity-sensitive). Because stellar activity can affect all the lines in the spectrum, the sparsity assumption is used to find the subset of lines that are most sensitive to activity.

We apply this method to study HARPS spectra of α Centauri B using the S-index and find many common activity-sensitive lines between our method and the method proposed by Wise et al. (2018). Our method is not limited to the use of the five activity indicators mentioned above. Other non-spectral indicators, like simultaneous photometric measurements, can be used to identify activity-sensitive lines as well, as long as the nonspectral indicators represent some manifestations of stellar activity in future studies. Moreover, our method can be used in other important applications. For example, the results of the proposed method can be used to identify differences in activitysensitive lines between stars with different ages and compositions, or to provide more precise RV measurements by removing activity-sensitive lines when constructing the cross-correlation function.

2 Method

Our sparse linear regression model is defined as:

$$Y_t = X_t\beta + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2 I_{n_t}), \tag{1}$$

where each Y_t is a $n_t \times 1$ vector of an activity indicator, $t = 1, \ldots, T$, T is the total number of days in the dataset and $\sum_{t=1}^{T} n_t$ is the total number of spectra. Let $Y = (Y_1', \ldots, Y_T')'$ be a $\sum_{t=1}^{T} n_t \times 1$ vector, which is normalized by removing the mean (centered) and then dividing the standard derivation (scaled). X_t is an $n_t \times p$ matrix containing the normalized flux of a given spectrum, which is also centered and then scaled, and p is the total number of pixels. β is an unknown p-dimensional sparse vector which needs to be estimated. Each coordinate in β, denoted by β_j, is the coefficient of the j-th pixel. The ϵ_ts are independent random errors, each of which follows a multivariate Gaussian distribution with variance $\sigma_t^2 I_{n_t}$, where I_{n_t} stands for an $n_t \times n_t$ identity matrix.

We assume β is sparse—that is, only a limited number pixels are highly sensitive to stellar activity. For each β_j, the prior is a probability density constructed from a linear combination of two distributions: a Laplace distribution with a relatively large variance if a pixel j is selected as activity-sensitive (non-zero), and a Laplace distribution with a very small variance if j is insensitive (zero). More explicitly, the spike-and-slab lasso prior (Rockova and George, 2018) is a product of p mixtures of two probability densities, defined as

$$(\beta|\gamma) \sim \prod_{j=1}^{p} [\gamma_j \psi(\beta_j|v_1) + (1-\gamma_j)\psi(\beta_j|v_0)], \tag{2}$$

where $0 < v_1 \ll v_0$, and $\psi(\beta_j|v) = \frac{v}{2} e^{-v|\beta_j|}$ is the Laplace density with mean 0 and variance $2v^{-2}$.

The parameter $\gamma_j \in \{0, 1\}$ is the weight, which determines whether the corresponding parameter β_j should be classified as zero or non-zero. When $\gamma_j = 1$, the density $\psi(\beta_j|v_1)$ is used for β_j. When $\gamma_j = 0$, the density $\psi(\beta_j|v_0)$ is used instead to force β_j to be close to 0 as v_0 is very large. Since γ_j is unknown, for each γ_j, we choose a prior as follows:

$$(\gamma_j|\theta) \sim \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}. \tag{3}$$

where $\theta \in (0, 1)$ is a hyperparameter.

We adopt a computationally feasible Expectation-Maximization (EM) algorithm to compute the posterior. The EM algorithm is an optimization algorithm which contains two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, γ is treated as a latent variable, and the expectation is taken with respect to the latent variable. In the M-step, the optimal values for the remaining parameters are solved. The two steps are repeated with many iterations until the optimal solution which maximizes the log-posterior is obtained.

3 Results

In this section, we highlight three major results found by using the method. First, we found that not only the pixels around the center of the line are active, but some pixels at the left side of the line are also variable (see Figure 1). It is interesting to observe that the pixels at the left side of the line have negative values while the pixels around the center of the line have positive values. The positive-valued pixels near the center may indicate heating in the chromosphere causing emission in the core of Hα, but a physical explanation for the negative pixels values on the left side of Hα is beyond the scope of this work.

Second, a list of the top 40 lines identified for the five indices is provided. The result for using the S-index in given in Table 1. In this table, a window size of ± 1 Å is used to determine if selected wavelengths belong to the same line, and the wavelength of the pixel with the highest magnitude regression coefficient is given.

Last, we calculated the percentage of overlapping lines obtained by using the five different activity indicators—S-index, NaD, Hα, the bisector inverse slope (BIS), and the full width at half maximum (FWHM)—and found that NaD index and S-index had 75% active lines that are overlapping (see Table 2). This result is expected by astronomers as emission in the cores of the Ca II H & K lines and the NaD lines are both probing activity in the lower chromosphere (Thatcher et al. 1991). BIS and FWHM are both expected to probe granulation (Gray 2009) and we found that 71% of lines selected by these two activity indicators are the same.

Table 1: Wavelengths obtained from S-index.

Wavelength	Species (VALD Depth)	Note
4427.33	Fe I (0.97), Fe I (0.82), V I (0.11)	1
5110.41	Fe I (0.93), Fe I (0.72), Fe I (0.21), Fe I (0.16)	1
4461.66	Fe I (0.96)	1
5429.70	Fe I (0.92)	1
6562.79	H I (0.45) (H α), Co I (0.24)	1
4375.94	Fe I (0.97), Fe I (0.8), Ce II (0.14)	1
4340.31	Fe I (0.54), H I (0.48) (H γ), Fe I (0.43), Fe I (0.21), Fe I (0.18)	
5895.81	Na I (0.9)	1
5397.13	Fe I (0.93), Fe I (0.52), Ti I (0.34)	1
6439.00	Ca I (0.77)	
5204.46	Cr I (0.92), Fe I (0.9)	
4957.57	Fe I (0.94), Fe I (0.91), Dy II (0.11)	
6122.15	Ca I (0.8)	
5012.08	Fe I (0.93), Fe I (0.67)	1
6496.94	Ba II (0.67), Ba II (0.47), Ba II (0.42), Ba II (0.39), Ba II (0.22)	
5432.56	Mn I (0.74)	1
5098.76	Fe I (0.87), Fe I (0.74)	
6162.12	Ca I (0.82)	
6318.12	Fe I (0.72), Ca I (0.42), Ti I (0.14)	
5434.52	Fe I (0.92)	1
4626.23	Cr I (0.9), Fe I (0.13)	
5226.82	Fe I (0.88), Fe I (0.36)	1
5225.48	Fe I (0.87)	
5497.59	Fe I (0.9), Y II (0.14)	1
5191.42	Fe I (0.88), Nd II (0.18)	
5269.54	Fe I (0.94)	1
5506.80	Fe I (0.9)	1
4534.01	Ti II (0.9), Fe I (0.51), Co I (0.45), Cr I (0.15)	
6462.49	Ca I (0.75)	
5409.86	Cr I (0.89), Ti I (0.41)	
5365.36	Fe I (0.75)	
5446.94	Fe I (0.92), Fe I (0.81)	
6219.23	Fe I (0.75)	
5405.80	Fe I (0.93)	1
5262.30	Ca I (0.74)	
5051.71	Fe I (0.92)	1
5281.74	Fe I (0.86), Ni I (0.22), Fe I (0.12)	
4861.54	H I (0.49), Cr I (0.42) (H β)	1
4981.82	Ti I (0.91)	
6495.08	Fe I (0.82)	

The 1 indicates the wavelengths also selected by WDBP18's method.

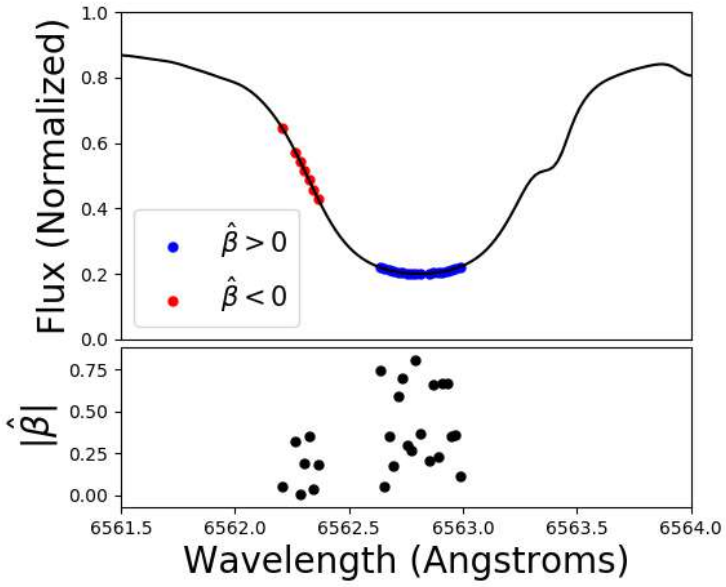


Figure 1: Selected wavelengths (the upper part) and regression coefficients (the lower part) in Hα using S-index as the response variable.

Table 2: Fractional overlap between selected lines for different indicators with the total number of **lines** selected on the diagonal. The tildes indicate our "within 1 Å" rule for grouping pixels into lines may be slightly inaccurate.

	Hα	FWHM	BIS	NaD	S
Hα	∼100	0.54	0.22	0.36	0.58
FWHM		∼440	0.71	0.72	0.63
BIS			∼63	0.35	0.76
NaD				∼67	0.75
S					∼273

Bibliographie

Gray F. G. (2009). The third signature of stellar granulation. *The Astrophysical Journal*. 697:1032–1043

Jurgenson, C., D. A. Fischer, T. McCraken, and et al. (2016). EXPRES: a next generation RV spectrograph in the search for earth-like worlds, *Proc. SPIE 9908, Ground-based and Airborne Instrumentation for Astronomy VI*, 99086T.

Ning, B., A. Wise, J. Cisewski-Kehe, S. Dodson-Robinson, and D. A. Fischer (2019). Identifying activity-sensitive spectral lines: A Bayesian variable selection approach. *The Astronomical Journal*, 158:15 pages

Rockova, V. and E. George (2018). The spike-and-slab lasso. *Journal of American Statistical Association*, 113: 431-444.

Thatcher, J. D., R. D. Robinson, and D. E. Rees (1991). The chromospheres of late-type stars - I. ϵ Eridani as a test case of multiline modelling. *Monthly Notices of the Royal Astronomical Society*, 250(1):14?23

Wise, A. W., S. E. Dodson-Robinson, K. Bevenour, and A. Provini (2018). New methods for finding activity-sensitive spectral lines: Combined visual identification and an automated pipeline find a set of 40 activity indicators. *The Astronomical Journal*, 156(180): 11 pages

STACKING PREDICTION FOR A MULTICLASS

Hicham Noçairi[1] & Fleur Tourneix[1]

[1] *L'Oréal – 1 Av. Eugène Schueller Aulnay-sous-Bois E-mail hicham.nocairi@rd.loreal.com*
[1] *L'Oréal – 1 Av. Eugène Schueller Aulnay-sous-Bois E-mail fleur.tourneix@rd.loreal.com*

Abstract. A large number of supervised classification models have been proposed in the literature. In order to avoid any bias induced by the use of one single statistical approach, they are combined through a specific "stacking" meta-model.
To deal with the case of a multiclass outcome and of categorical predictors, we introduce several improvements to stacking: combining models is done through PLS-DA instead of OLS due to the strong correlation between predictions, and a specific methodology is developed for the case of a small number of observations, using repeated sub-sampling for variables selection. Five very different models (Boosting, Bagging, Random Forest, ANN, SVM, KKNN and Sparse PLS-DA) are combined through this improved stacking, and applied in the context of the development of alternative strategies for safety evaluation where multiple *in vitro*, *in silico* and physicochemical parameters are used to classify substances in three classes : "ES = Extreme Strong", "WM = Weak Moderate" and "NS = Non-Sensitizer". Results show that stacking meta-models have better performances than each of the seven models taken separately, and furthermore, stacking provides a better balance sensitivity and specificity.

Keywords. Stacking meta-model, Multiclass outcome, Prediction, Sparse-PLSDA, Boosting, Bagging, Random Forest Artificial network, SVM, CART.

Bibliographie

Bühlmann, P., & Hothorn, T. (2007), Boosting Algorithms: Regularization, Prediction and Model Fitting. Institute of Mathematical Statistics, *Statistical Science*, Vol 22, 7, 477-505.
Chung, D. & Keles, S. (2010), Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, Vol. 9, Article 17.
Becker, N., Werft, W., & Benner, A. (2010), penalizedSVM: Feature Selection SVM using penalty functions, R package version 1.1. http://CRAN.Rproject.org/package=penalizedSVM.
Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.
Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). Classification and Regression Trees. Boca Raton, FL: CRC press.
Zhang, G. P., & Patuwo, M. Y. H. (1998). Forecasting with artificial neural networks: The state of the art. International Archives of Photogrammetry and Remote Sensing, 1(14), 35–62.
LEO BBEIMAN, (1996), Bagging Predictors, *Machine Learning*, 24, 123-140.
Noçairi H., Gomes C., Thomas M., & Saporta G., (2016), Improving-stacking methodology for combining classifiers: applications to cosmetic industry. Electronic Journal of Applied Statistical Analysis, Vol. 09, Issue 02, 340-361.

Analyse de données d'épidémie de malaria par un modèle de fragilité multivarié à corrélations spatiales

Ajmal Oodally [1] & Klara Goethals [2] & Estelle Kuhn [3] & Luc Duchateau [4]

[1] *ajmal.oodally@inrae.fr*
[2] *klara.goethals@ugent.be*
[3] *estelle.kuhn@inrae.fr*
[4] *luc.duchateau@ugent.be*

Résumé La malaria est une maladie avec un taux de mortalité qui reste élevé en Afrique sub-saharienne. La transmission se fait via un moustique, dont le développement et la reproduction sont favorisés par la présence de plans d'eau. Afin d'étudier l'influence des plans d'eau sur le taux de transmission de la malaria, un riche jeu de données a été constitué dans le secteur du barrage hydroélectrique de Gilgel Gibe dans le sud-ouest de l'Éthiopie. Il est constitué de temps d'infection par la malaria d'enfants répartis en villages, ainsi que de nombreuses covariables. Ces données ont déjà été analysées par un modèle de fragilité structuré selon les villages, incluant la distance entre l'enfant et le barrage comme covariable. Cependant, la proximité entre les enfants n'est pas prise en compte. Afin de mieux prendre en compte cette spécificité liées aux données, nous proposons un modèle de fragilité avec une structure de corrélation spatiale. Les paramètres du modèle sont estimés en maximisant la vraisemblance observée via un algorithme de type Expectation Maximization stochastique. La performance de l'estimateur est évaluée sur des données simulées et des données réelles.

Mots-clés. Modèle de fragilité multivarié, corrélations spatiales, algorithme Expectation Maximization stochastique, incidence de la malaria

Summary

Malaria remains a disease with high morbidity and mortality in Subsaharan Africa, and more specifically in Ethiopia. The mosquito being the vector of this disease, the presence of water bodies, favoring the reproduction and breeding of the mosquito, strongly influences the rate of transmission. A rich dataset has been put together in the area of the Gilgel Gibe hydroelectric dam in south-western Ethiopia, including the malaria infection times of

2037 children situated in different villages as well as many covariates. The data has been analyzed by frailty models introducing the village as cluster to accommodate for the correlation in the data and distance from the dam as main risk factor. However, proximity between the children is not taken into account. In order to consider this specificity in the data, we propose a frailty model with a spatial correlation structure. The parameters of the model are estimated by maximizing the observed likelihood via a stochastic Expectation Maximization algorithm. Parameter estimation is done on the malaria dataset and simulated data to assess the performance of the estimator.

Keywords. Multivariate frality model, spatial correlation, stochastic Expectation Maximization algorithm, malaria incidence

1 Contexte, problématique et données

La malaria est une maladie avec un taux de mortalité qui reste élevé en Afrique sub-saharienne. Le parasite de la malaria est transmis d'homme à homme par le moustique anophèle. Ce moustique est fortement dépendant de l'eau à tous les stades de son développement (œuf, larve, nymphe) et pour sa reproduction. De fait, la présence de plans d'eau, comme ceux liés à la construction de barrages hydrauliques, peut potentiellement avoir un fort impact sur l'incidence de la malaria. En 2003, le barrage hydroélectrique de Gilgel Gibe a été construit dans le sud-ouest de l'Éthiopie. Afin d'étudier l'effet du réservoir d'eau sur la propagation de la malaria, 16 villages à différentes distances du barrage, variant entre 0,26 et 9,05 km, ont été sélectionnés. Au total, 2037 enfants de moins de 10 ans ont fait l'objet d'un suivi hebdomadaire entre juillet 2008 et juin 2010. Les temps d'infection par la malaria ont été observés sur ces enfants avec censure, ainsi que de nombreuses covariables descriptives. Pour plus d'information sur ces données, nous renvoyons aux articles de Yewhalaw et al (2009), Getachew et al (2013). L'hypothèse principale de ces travaux était que plus on s'éloigne du barrage, plus l'incidence de la malaria diminue. Les données ont été analysées via des modèles d'analyse de survie pour mieux quantifier cet effet.

Les modèles de fragilité introduits par Vaupel et al (1979) sont une extension du modèle de Cox (1972) qui modélise le risque de survenue d'un événement comme produit d'une fonction de risque de base et d'une fonction des covariables. Ces modèles permettent en outre de prendre en compte l'hétérogénéité présente dans les données via des effets aléatoires non ob-

servés. Ainsi un modèle de fragilité incluant l'effet "village" comme effet aléatoire et la distance entre l'individu et le barrage comme facteur de risque principal a été utilisé par Getachew et al (2013). Cependant, cette modélisation soulève deux remarques. Premièrement, les résultats du modèle de fragilité ne sont pas fiables lorsqu'il existe une forte corrélation entre la covariable, ici la distance au barrage, et le groupe, ici le village, ce qui est le cas dans ce jeu de données. Deuxièmement, le village est une structure administrative qui ne rend pas forcément compte de la proximité géographique entre individus.

Une alternative consiste à prendre en compte les distances entre individus qui jouent un rôle important dans la transmission de la malaria. Un individu infecté pose un risque à tous ceux qui sont dans son voisinage puisque tout moustique qui le pique peut ensuite transmettre la maladie. Par conséquent, nous proposons un modèle de fragilité avec une structure de corrélation spatiale au niveau de l'individu.

2 Modèle de fragilité multivarié à corrélation spatiale

Usuellement, les modèles de fragilité spatiaux modélisent les corrélations entre groupes (cf. Li et Ryan (2002)). Nous proposons un modèle de fragilité multivarié à corrélations spatiales au niveau de l'individu. On considère une population de N individus. Pour $1 \leq i \leq N$, le temps de survenue de l'infection et le temps de censure pour l'individu i sont modélisés par des variables aléatoires notées T_i et C_i respectivement. On observe alors pour $1 \leq i \leq N$ le temps censuré à droite et l'indicateur de censure notés respectivement X_i et Δ_i et définis par $X_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{T_i \leq C_i}$. Dans la suite, on note $\mathbf{X} = (X_i)_{1 \leq i \leq N}$ et $\mathbf{\Delta} = (\Delta_i)_{1 \leq i \leq N}$. Le modèle de fragilité multivarié à corrélations spatiales est défini pour $1 \leq i \leq N$ par :

$$h_i(t|b_i) = h_0(t)\ \exp(Z_i^t \beta + b_i) \quad i = 1, \ldots, N \text{ où } \ b = (b_i)_{1 \leq i \leq N} \sim \mathcal{N}_N(0, \Gamma)$$

où $h_i(t|b_i)$ désigne le risque instantané de survenue de l'événement pour l'individu i au temps t, $h_0(t)$ le risque de base au temps t, b_i le vecteur de fragilité de l'individu i, β le vecteur des paramètres de régression inconnu, Z_i le vecteur de covariables associées à l'individu i. Le vecteur de fragilité b suit une distribution normale multivariée centrée avec une matrice de covariance

notée $\Gamma = \sigma^2\Sigma(\rho)$ où σ^2 est un facteur d'échelle et $\Sigma(\rho)$ est la matrice de corrélation structurée paramétrée par $\rho > 0$. Nous considérons les deux structures suivantes :

$$\Sigma_1(\rho) = \exp(-\rho D) \quad \text{et} \quad \Sigma_2(\rho) = \frac{1}{1+D^\rho} \tag{1}$$

où $D = (d_{ij}) \in R^{N\times N}$ est telle que la composante d_{ij} correspond à la distance entre l'individu i et l'individu j pour $i \neq j$ et $d_{ii} = 0$ par convention.

On suppose par ailleurs que la fonction de risque de base h_0 est paramétrique constante par morceaux pour prendre en compte l'effet saisonnier de l'incidence. La fonction de risque de base est définie par $h_0(t) = h_m$ pour $t \in [\tau_{m-1}, \tau_m[$ pour $m \in [1, M]$ où $(\tau_m)_{m\in[1,M]}$ est une suite strictement croissante et $\tau_0 = 0$. Le modèle s'écrit alors :

$$h(t|b_i) = \sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(Z_i^t\beta + b_i) \tag{2}$$

Les paramètres à estimer sont $\theta = ((h_m)_{1\leq m\leq M}, \beta, \sigma^2, \rho)$.

3 Estimation des paramètres

Nous estimons les paramètres du modèle par maximum de vraisemblance. La log-vraisemblance complète des données s'écrit :

$$\begin{aligned}\log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b) \;=\; & \sum_{m=1}^{M}\sum_{i=1}^{N}\Bigg(\Delta_i\Big(\log(h_m)\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_i) + Z_i^t\beta + b_i\Big) \\ & -H(X_i)\exp(Z_i^t\beta + b_i)\Bigg) + \sum_{i=1}^{N}\log f_\Gamma(b_i)\end{aligned}$$

où le hasard cumulé est noté $H(X_i) = \sum_{l=1}^{M} h_l(\tau_l - \tau_{l-1})\mathbb{1}_{\tau_l\leq X_i} + \sum_{l=1}^{M}(X_i - \tau_{l-1})h_l\mathbb{1}_{\tau_{l-1}\leq X_i<\tau_l}$. La vraisemblance observée est obtenue en intégrant la vraisemblance complète par rapport au vecteur de fragilité b :

$$L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \int L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b)db$$

On définit $\hat{\theta}$ l'estimateur du maximum de la vraisemblance observée par $\hat{\theta} = \text{argmax}\, L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta})$. Le calcul de cet estimateur ne peut en général pas

se faire directement, en particulier lorsque la vraisemblance observée n'admet pas de forme analytique, ce qui est le cas dans le modèle de fragilité défini dans la section 2. En pratique, nous calculons la valeur de l'estimateur via un algorithme itératif stochastique de type Expectation Maximization.

4 Algorithme stochastique d'estimation

On applique l'algorithme MCMC-SAEM introduit par Kuhn et Lavielle (2004) qui combine une méthode de Monte Carlo Markov Chain à l'algorithme d'approximation stochastique de EM. Chaque itération de l'algorithme se compose de trois étapes. A l'itération k :

S-step : une réalisation b^k du vecteur de fragilité non observé est simulé selon le noyau de transition d'une chaîne de Markov convergente $\Pi_{\theta_{k-1}}$ ayant comme distribution stationnaire la distribution conditionnelle du vecteur de fragilité notée $\pi_{\theta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta})$.

$$\pi_{\theta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta}) = \frac{L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b)}{L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta})}$$

SA-step : On effectue une approximation stochastique sur la log-vraisemblance complète :

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b^k) - Q_{k-1}(\theta))$$

où la suite (γ_k) vérifie $0 \leq \gamma_k \leq 1, \sum \gamma_k = +\infty, \sum \gamma_k^2 < +\infty$.

M-step : On met à jour les paramètres selon : $\theta_k = \underset{\theta}{\text{argmax}}\ Q_k(\theta)$

Les quantités Q_0 et θ_0 sont initialisées arbitrairement. Sous des hypothèse de régularité du modèle de fragilité et sous des hypothèses assurant l'ergodicité de la chaine de Markov, la suite $(\theta_k)_k$ générée par l'algorithme décrit ci-dessus converge presque sûrement vers un point critique de la vraisemblance observée (cf. Kuhn et Lavielle (2004)).

5 Expériences numériques

On analyse les temps de survenue de la malaria chez 2037 enfants. Les quatres covariables considérées sont l'âge, le sexe, la structure du toit et la distance

Table 1 – Estimations des paramètres et écart type entre parenthèses

β_{sex}	β_{age}	β_{d}	β_{roof}	$(h_1, h_2, h_3, h_4, h_5, h_6) \times 10^{-4}$	σ^2	ρ
-0.0391	-0.0061	0.1057	0.0260	(5.40, 14.3, 3.98, 6.88, 2.42, 2.71)	0.364	0.794
(0.0659)	(0.0201)	(0.140)	(0.0342)	(0.48,0.97,0.22,0.49,0.11,0.18)	(0.088)	(0.11)

entre l'enfant et le barrage. Du fait de la grande dimension du vecteur de fragilité et de l'hétérogénéité de ses composantes, l'étape de simulation est délicate et nécessite l'utilisation de techniques adaptées à ce contexte. Ainsi la simulation du vecteur b^k à l'itération k de l'algorithme se fait via un algorithme de Gibbs hybride adaptatif (cf. Atchadé et Rosenthal (2005)), assurant un taux d'acceptation homogène selon toutes les composantes.

Nous avons testé la méthode d'estimation sur des données simulées selon le modèle défini dans l'équation (2) et avons obtenu de très bons résultats. Pour l'analyse des données de malaria, nous comparons plusieurs modèles avec différentes fonctions de risque de base h_0 et les structures de corrélation présentées dans la section 2. Les modèles que nous comparons ont le même nombre de paramètres et nous présentons donc les résultats du modèle qui maximise la log-vraisemblance dans le tableau 1. L'estimation de h_2 correspond à la plus forte période de pluie et est associée à une plus grande incidence. La présence de plans d'eau est plus importante pendant la grosse saison de pluie et favorise la reproduction des moustiques et contribue donc à propager plus facilement la malaria. Aussi, nous illustrons graphiquement ce que représente l'estimation du paramètre de corrélation ρ dans la figure 1. Cette estimation semble cohérente avec la distance maximum théorique que parcours le moustique.

Bibliographie

Cox, D.R. (1972). Regression Models and Life-Tables, *Journal of the Royal Statistical Society*, 34, pp. 187-220. Vaupel, J., Manton, K. et Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, 16, pp. 439-454. Getachew, Y., Janssen, P., Yewhalaw, D., Speybroeck N. et Duchateau L. (2013). Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models, *Statistics in Medicine*, 32, pp. 3224-3233. Yewhalaw, D., Worku L.,

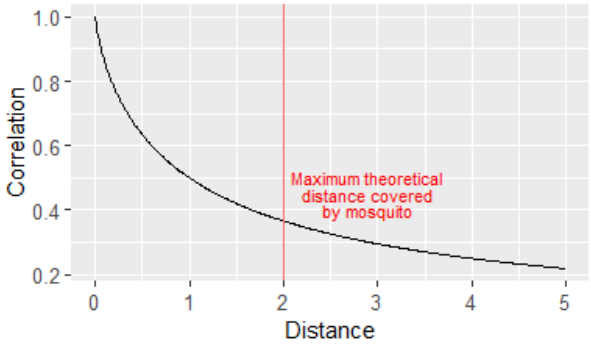


Figure 1 – Représentation graphique de la corrélation en fonction de la distance basée sur l'estimation de $\hat{\rho} = 0.794$

Van Bortel W., GebreSelassie S., Kloos H., Duchateau L. et Speybroeck N. (2009). Malaria and water resource development: the case of Gilgel-Gibe hydroelectric dam in Ethiopia, *Malaria Journal*, 8, 21. Dempster, A. P., Laird N. M. et Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1, pp. 1-38. Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure, *ESAIM: Probability and Statistics*, 8, pp. 115-131. Atchadé, Y. et Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms, *Bernoulli*, 11. Li, Y. et Ryan, L. (2002). Modeling Spatial Survival Data Using Semiparametric Frailty Models, *Biometrics*, 58, pp. 287-297.

Transfer learning pour la régression linéaire & test de gain

David Obst [1] & Badih Ghattas [2] & Jairo Cugliari [3] & Georges Oppenheim [4]

[1] *EDF R&D, Palaiseau, France / I2M, Aix-Marseille Université, France; david.obst@edf.fr*
[2] *I2M, Aix-Marseille Université, France; badih.ghattas@univ-amu.fr*
[3] *Laboratoire ERIC, Université de Lyon 2, France; jairo.cugliari@univ-lyon2.fr*
[4] *LAMA, Université Paris-Est, Champs-sur-Marne, France; georges.oppenheim@gmail.com*

Résumé. Le Transfer Learning, parfois aussi appelé transfert de connaissances, a pour objectif de réutiliser des connaissances d'une tâche source afin d'améliorer les performances sur une autre cible qui lui est semblable. De nombreux résultats empiriques attestent de l'utilité du transfert, mais peu de résultats théoriques existent pour les problèmes de régression. Dans ce papier nous introduisons un cadre théorique pour le transfert dans le cadre de la régression linéaire. Il est montré que la qualité du transfert pour un nouveau vecteur x dépend de sa représentation sur une certaine base propre faisant intervenir les paramètres du problème. Par ailleurs nous aboutissons à la construction d'un test statistique permettant de prévoir si un modèle amélioré par transfert aura une erreur de prévision plus basse que le modèle cible de base pour une nouvelle observation x. L'efficacité du test est illustrée sur des données synthétiques ainsi que des véritables données de consommation électrique.

Mots-clés. Régression linéaire, Transfer learning, Test statistique, Fine-tuning, Théorie du transfert

Abstract. Transfer learning, also referred as knowledge transfer, aims at reusing knowledge from a source dataset to a similar target one. While many empirical studies illustrate the benefits of transfer learning, few theoretical results are established especially for regression problems. In this paper a theoretical framework for the problem of parameter transfer for the linear model is proposed. It is shown that the quality of transfer for a new input vector x depends on its representation in an eigenbasis involving the parameters of the problem. Furthermore a statistical test is constructed to predict whether a fine-tuned model has a lower prediction quadratic risk than the base target model for an unobserved sample. Efficiency of the test is illustrated on synthetic data as well as real electricity consumption data.

Keywords. Linear regression, Transfer learning, Statistical test, Fine-tuning, Transfer theory

1 Introduction

On considère la situation où l'on cherche à réaliser des prévisions pour une tâche cible $\mathcal{T}$ pour laquelle le nombre d'échantillons d'entraînement est limité. Néanmoins nous disposons également d'une tâche source $\mathcal{S}$ qui lui est apparentée et pour laquelle le nombre d'échantillon est plus conséquent. Notre objectif est de tirer profit des données de $\mathcal{S}$ afin d'améliorer les performances de prévision sur $\mathcal{T}$ (Weiss et al.; 2016). Pan et Yang (2016) définissent quatre catégories du transfert: de paramètres, d'instances, de features et de relations. Nous considérons dans notre papier le transfert de paramètre dans le cadre de la régression linéaire. Le transfert pour la régression linéaire a déjà été étudié a plusieurs reprises. Bouveyron et Jacques (2010) proposent une approche où l'estimation des paramètres cible est obtenue par une transformation linéaire sous contraintes du paramètre source. Récemment Dar et Baraniuk (2020) ont étudié le transfert de paramètres dans le cadre restreint où les variables explicatives sont générées de manière i.i.d. selon une loi normale $\mathcal{N}(0, I_D)$ où I_D la matrice identité de taille D. Ils montrent l'existence d'un phénomène de "double double descente" où le transfert est bénéfique dans les cas où les tâches sont sous ou sur-paramétrées. Chen et al. (2015) ont suggéré une autre approche de transfert de paramètres pour le modèle linéaire obtenu par une combinaison convexe ou matricielle des estimations source et cible. Dans le cas de certaines hypothèses sur les données, cette combinaison a la garantie d'être meilleure que l'estimateur cible seul. Néanmoins les papiers cités précédemment ont tous les point commun de ne pas étudier une approche notable du transfer learning, à savoir le fine-tuning. Par ailleurs les hypothèses faites sur les données sont souvent restrictives et rarement vérifiées en pratique.

2 Gain du transfert

On considère deux tâches de régression linéaire indépendantes. Soit (X_S, Y_S) le jeu de données source de taille N_S, où X_S est la matrice du plan d'expérience de dimension $N_S \times D$, $Y_S = X_S\beta_S + \varepsilon_S$ avec $\beta_S \in \mathbb{R}^D$ le coefficient à estimer et $\varepsilon_S \sim \mathcal{N}(0, \sigma_S^2 I_{N_S})$ un bruit i.i.d. Similairement nous avons le jeu cible (X_T, Y_T) de taille N_T (avec $N_T \ll N_S$) où $\beta_T \in \mathbb{R}^D$, $X_T \in \mathbb{R}^{N_T \times D}$, $Y_T = X_T\beta_T + \varepsilon_T$ et $\varepsilon_T \sim \mathcal{N}(0, \sigma_T^2 I_{N_T})$.

2.1 Fine-tuning

En notant $\Sigma_\nu = X_\nu^\top X_\nu$ et en supposant les X_ν de rang colonne plein, l'estimateur habituel obtenu en minimisant l'objectif des moindres carrés $J_\nu(\beta) = \frac{1}{2}\|Y_\nu - X_\nu\beta\|^2$ est $\hat{\beta}_\nu = \Sigma_\nu^{-1} X_\nu^\top Y_\nu$ avec $\nu \in \{S, T\}$. Néanmoins si le nombre d'échantillons cible N_T est trop faible, l'estimateur $\hat{\beta}_T$ sera de mauvaise qualité. L'idée est donc de partir de l'estimateur $\hat{\beta}_S$ qui fournit une base solide et de l'ajuster sur la tâche cible. C'est dans cette optique qu'une approche de fine-tuning à base de descente de gradient à pas fixe α sur la fonction

objectif $J_T(\beta)$ et partant de $\hat{\beta}_S$ a été retenue. On peut prouver alors que l'estimateur $\hat{\beta}_k$ obtenu au bout de k itérations est donné par (1):

$$\hat{\beta}_k = A^k \hat{\beta}_S + (I_D - A^k)\hat{\beta}_T \tag{1}$$

où $A = I_D - \alpha\Sigma_T$. Il s'agit alors d'une forme particulière des estimateurs de la forme $\hat{\beta}(W) = W\hat{\beta}_S + (I_D - W)\hat{\beta}_T$ mentionnés dans Chen et al. (2015).

2.2 Formulation du gain

La principale difficulté dans la littérature est de définir la notion de transfert bénéfique et négatif. Dans le cadre d'un problème de prévision, il semble naturel de dire que le transfert est bénéfique quand l'erreur de prévision obtenue avec le modèle enrichit par transfert est inférieure à celle que l'on aurait uniquement avec ce qui est appris en cible pour une nouvelle observation. Nous définissons donc le *gain* $\Delta\mathcal{R}_k(x)$ quantifiant l'apport du transfert pour un nouvel échantillon cible $(x, y) \in \mathbb{R}^D \times \mathbb{R}$ par:

$$\Delta\mathcal{R}_k(x) = \mathbb{E}[(y - x^\top \hat{\beta}_T)^2] - \mathbb{E}[(y - x^\top \hat{\beta}_k)^2] \tag{2}$$

On peut alors prouver que le gain (2) prend la forme particulière:

$$\Delta\mathcal{R}_k(x) = x^\top H_k x \quad \text{with} \quad H_k = \sigma_T^2(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k) - \sigma_S^2 A^k \Sigma_S^{-1} A^k - A^k B A^k \tag{3}$$

où $\Omega_k = \frac{1}{\alpha}\Sigma_T^{-1}(I_D - A^k)$ et $B = (\beta_T - \beta_S)(\beta_T - \beta_S)^\top$. Ainsi quand $\Delta\mathcal{R}_k(x) > 0$, le transfert sera bénéfique pour l'échantillon (x, y). On voit donc que le signe du gain va dépendre de la représentation de x sur la base propre associée à la matrice H_k, puisque le gain sera positif pour les vecteurs dans la somme des espaces propres de H_k associés aux valeurs propres positives. En fonction de la représentation de x sur ces espaces, il peut donc être judicieux ou non d'utiliser le modèle fine-tuned. Un problème immédiat est donc de savoir pour quels x utiliser le nouveau modèle. Par ailleurs le gain introduit par (2) est généralisable à d'autres estimateurs tels que ceux de Chen et al. (2015).

3 Test

Il serait naturel de chercher un estimateur de $x^\top H_k x$. Néanmoins après plusieurs essais, une telle estimation est souvent de mauvaise qualité et inutilisable.

3.1 Formulation du test

C'est pour cela que nous considérons plutôt le problème de prévoir si le gain sera positif ou non pour un x donné, ce qui peut se résumer sous la forme d'un test d'hypothèse.

On peut alors montrer que le test donné par (4) est approximativement de niveau a pour tester l'hypothèse nulle $H_0 : \{\Delta\mathcal{R}_k(x) \leq 0\}$ contre l'alternative $H_1 : \{\Delta\mathcal{R}_k(x) > 0\}$:

$$\mathbb{1}\Big(\frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{x^\top(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k)x - \rho^2\left\|A^k x\right\|^2}{x^\top A^k\Sigma_S^{-1}A^k x} > q^{1-a}\Big) \tag{4}$$

où les $\hat{\sigma}_\nu^2$ désignent les estimateurs habituels de la variance q^{1-a} est le quantile d'ordre $1-a$ de la distribution $\mathcal{F}(N_T - D, N_S - D)$ de Fisher et $\rho \geq \|\beta_T - \beta_S\|/\sigma_T$.

3.2 Choix des hyperparamètres α, k et ρ

Trois quantités doivent être calibrées avant l'utilisation du test: le pas de gradient α, le nombre d'itérations k et enfin la constante ρ. Le choix du pas α n'est pas crucial puisqu'il peut être compensé par un nombre suffisant d'itérations k (tant qu'il est suffisamment petit pour assurer la convergence). Après de nombreux essais expérimentaux, $\alpha = \alpha^*/5$ avec $\alpha^* = 2/\big(\lambda_{\max}(\Sigma_T) + \lambda_{\min}(\Sigma_T)\big)$ (où $\lambda_{\max}(\Sigma_T)$ et $\lambda_{\min}(\Sigma_T)$ désignent respectivement les valeurs propres maximales et minimales de la matrice Σ_T). Le choix de k se fait en minimisant empiriquement le 2$^{\text{e}}$ terme à droite de (5) qui permet de maximiser le gain:

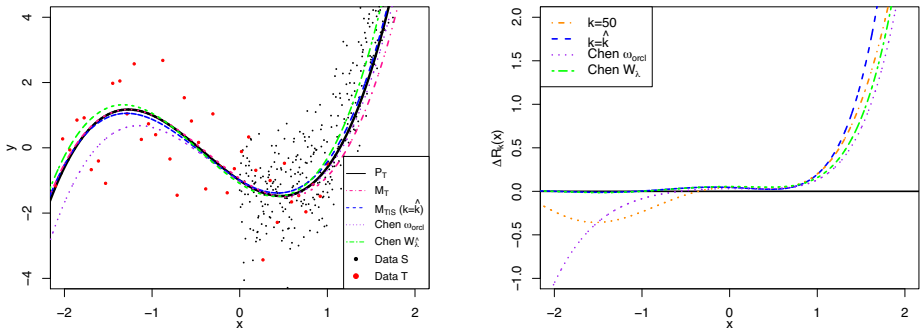
$$\Delta\mathcal{R}_k(x) = -2\sigma_T^2\, x^\top\Sigma_T^{-1}x\, D_{KL}(\mathcal{N}_k||\mathcal{N}_T) - \sigma_T^2\, x^\top\Sigma_T^{-1}x \ln\Big(\frac{x^\top V_k x}{\sigma_T^2\, x^\top\Sigma_T^{-1}x}\Big) \tag{5}$$

Enfin le choix de ρ se fait en évaluant la précision et le rappel du test (4) sur un jeu de validation. Les valeurs obtenues ainsi seront notées $\hat{k}$ et $\hat{\rho}$.

4 Simulations numériques

4.1 Données simulées

On cherche à estimer les coefficients d'un polynôme cible $P_T(x) = \beta_{T,0} + \beta_{T,1}x + \beta_{T,2}x^2 + \beta_{T,3}x^3$ où $\beta_T = (-1; -1.8; 1.2; 1)^\top$. On dispose de $N_T = 60$ observations i.i.d. $y_{T,i} = P_T(x_{T,i}) + \varepsilon_{T,i}$ avec $x_{T,i} \in [-3, 1]$. D'autre part nous disposons de $N_S = 600$ données source $y_{S,i} = P_S(x_{S,i}) + \varepsilon_{S,i}$ avec $x_{S,i} \in [0, 3]$ et $\sigma_S^2 = \sigma_T^2 = 1$. Les coefficients β_S sont obtenus en rajoutant un bruit $\mathcal{N}(0, 0.3^2)$ à ceux de β_T. Cet exemple a l'avantage d'être facilement visualisable et intuitif. Nous avons pris $\alpha = \alpha^*/5$ et k et ρ obtenus via les procédures décrites Section 3.2. Le vrai polynôme P_T ainsi que ses estimations obtenues via $\hat{\beta}_T$, $\hat{\beta}_k$ et deux estimateurs issus de Chen et al. (2015) sont représentés Fig. 1. On observe ainsi en 1.(a) que le fine-tuning permet d'avoir une meilleur estimation que le modèle cible pur, ce qui est confirmé par le gain en (b) notamment quand $x \geq 0$.
La p-value du test est tracée en fonction de x en Fig. 2. Elle est grande sur les plages de x où l'apport du transfert semblait moindre graphiquement. Typiquement on utiliserai $\hat{y}_T = x^\top\hat{\beta}_T$ quand elle est supérieure à 0.05, et $\hat{y}_k = x^\top\hat{\beta}_k$ sinon.



(a) Superposition de P_T et ses estimations.

(b) Gain théorique pour différents estimateurs.

Figure 1: Comparaison des estimations de P_T et du gain.

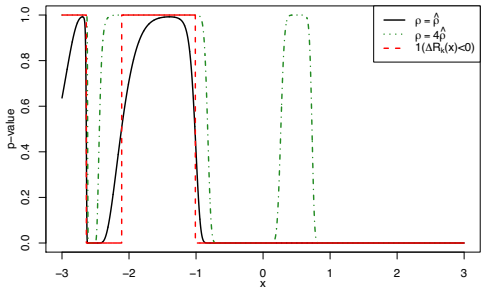


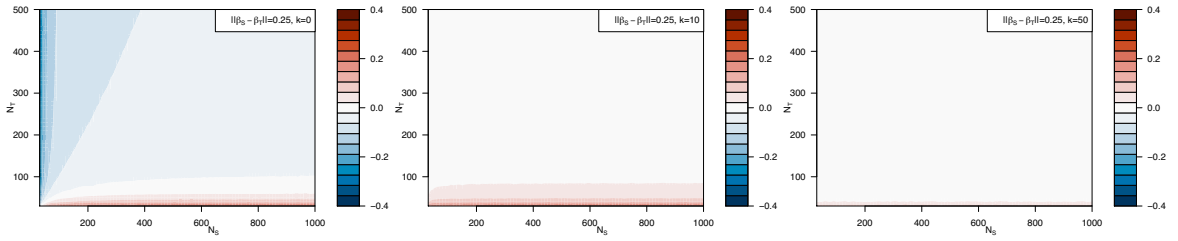
Figure 2: P-value du test en fonction de x ($k = \hat{k}$).

Les résultats obtenus sur ces données polynomiales ont été confortés sur des vraies données de consommation électrique issues de la compétition GEFCOM2012, illustrant l'intérêt pratique du test.

4.2 Gain en fonction des paramètres du problème

La formulation du gain (3) est qu'elle permet également d'interpréter les bénéfices du transfert en fonction des paramètres du problème N_ν, σ_ν^2, D et la distance $\|\beta_T - \beta_S\|$. Il est déjà clair que le gain diminue quand $\|\beta_T - \beta_S\|$ augmente (les tâches deviennent dissimilaires) ou quand σ_T^2 diminue (estimation en cible plus facile). Par ailleurs on s'attendrait que le gain soit élevé quand le nombre d'échantillons cible N_T est petit, le source N_S est grand. Ces hypothèses sont confirmées par la Figure 3 obtenues en moyennant d'un nombre important de simulations où les lignes de X_ν et x sont tirés i.i.d.

de $\mathcal{N}(0, I_D)$. Les zones rouges correspondant à un gain positif sont obtenues pour N_T petit. On remarque par ailleurs l'importance du choix de k, puisque $k = 10$ itérations de gradient permet de rendre le transfert exclusivement positif, mais un k trop important efface les bénéfices apportés par $\hat{\beta}_S$.



(a) $k = 0$ (b) $k = 10$ (c) $k = 50$

Figure 3: Phases de transfert en fonction de N_S, N_T and k.

Les résultats présentés dans ce papier ont été approfondis récemment. Ainsi la formulation du gain présenté ici donne également lieu à un phénomène analogue à celui mentionné par Dar et al. (2020), avec un gain qui tend vers l'infini quand $D \to N_T$ (qui ne nécessite donc pas l'hypothèse de rang plein des X_ν). Par ailleurs le rôle de α et de k peut être mieux compris en calculant $\mathbb{E}[\Delta\mathcal{R}_k(x)]$ dans le cadre des variables gaussiennes.

Bibliographie

Weiss, K., Khoshgoftaar, T. M., et Wang, D. (2016). *A survey of transfer learning.* Journal of Big data, 3(1), 1-40

Pan, S. J., et Yang, Q. (2009). *A survey on transfer learning.* IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359

Bouveyron, C., et Jacques, J. (2010). *Adaptive linear models for regression: improving prediction when population has changed.* Pattern Recognition Letters, 31(14), 2237-2247

Chen, A., Owen, A. B., and Shi, M. (2015). *Data enriched linear regression.* Electronic journal of statistics, 9(1), 1078-1112.

Dar, Y., and Baraniuk, R. G. (2020). *Double Double Descent: On Generalization Errors in Transfer Learning between Linear Regression Tasks.* arXiv preprint arXiv:2006.07002.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). *A theory of learning from different domains.* Machine learning, 79(1), 151-175.

Analyse statistique des signaux EEG/MEG pour la cognition et les interfaces cerveau-ordinateur

Théodore Papadopoulo

Université Côte d'Azur, INRIA, France.
Theodore.Papadopoulo@inria.fr

Résumé. L'électroencéphalographie et la magnétoencéphalographie sont deux méthodes qui permettent de mesurer des traces de l'activité électrique du cerveau. Ces mesures prennent la forme de séries temporelles vectorielles avec quelques dizaines à quelque centaines de capteurs qui fournissent des données à des fréquences de l'ordre de 256Hz à 2kHz. Les signaux mesurés ne sont pas forcément naturels et sont très bruités. Leur analyse (reconstruction de sources, classification, ...) est donc un véritable challenge et repose souvent sur une analyse de signaux moyennés sur un certain nombre de réalisations, ce qui masque souvent la variabilité intrinsèque de ces signaux. La prise en compte de ces variabilités intra ou inter-sujet est cruciale notamment dans les cas où moyenner plusieurs réalisations du signal n'est pas possible. C'est le cas des interfaces cerveau-ordinateur qui doivent pouvoir classifier les signaux cérébraux pour les transformer en actions en temps-réel. Mieux comprendre la variabilité des signaux EEG permettrait de s'affranchir (ou de réduire le temps qui y est consacré) les phases de calibration qui permettent au système d'apprendre les caractéristiques des sujets.

Mots-clés. Neurosciences, Interfaces Cerveau-Ordinateur, EEG/MEG, Classification,

Abstract. Electroencephalography and magnetoencephalography are two methods of measuring traces of the electrical activity in the brain. These measurements take the form of vector time series with a few tens to a few hundred of sensors which provide data at frequencies of the order of 256Hz to 2kHz. The measured signals are not necessarily natural and are very noisy. Their analysis (reconstruction of sources, classification, ...) is therefore a real challenge and is often based on an analysis of signals averaged over a number of realizations, which often masks the intrinsic variability of these signals. Taking account of these intra or inter-subject variabilities is crucial, in particular when averaging several realizations of the signal is not possible. This is the case with brain-computer interfaces which must be able to classify brain signals to transform them into actions in real-time. Better understanding the variability of EEG signals would allow to avoid (or reduce the time devoted to it) the calibration phases which allow the system to learn the specific characteristics of the subjects.

Keywords. Neurosciences, Brain-Computer Interfaces, EEG/MEG, Classification,

1 Statistical analysis of EEG/MEG signals for cognition and brain-computer interfaces

Electro-Encephalography (EEG) and Magneto-Encephalography (MEG) are two non-invasive techniques for measuring (part of) the electrical activity of the brain. Nowadays, EEG is relatively inexpensive and is commonly used to detect and qualify neural activity (epilepsy detection and characterisation, neural disorder qualification, BCI, ...). MEG and EEG can be measured simultaneously (M/EEG) and reveal complementary properties of the electrical fields.

The two techniques have temporal resolutions of about the millisecond, which is the typical granularity of the measurable electrical phenomena that arise in the brain. This high temporal resolution is what makes MEG and EEG attractive for the functional study of the brain. Their spatial resolution, on the contrary, is rather poor as only a few hundred of sensors can be placed around the head and acquired simultaneously (about 300-400 sensors for MEG and up to – but often much less – 256 sensors for EEG). Detecting and extracting meaningful information from the measurements is a difficult task, because of the low signal to noise ratio and the presence of ongoing cerebral activity (the notion of "noiseless signal" does not exist).

In most cases, only a subset of that piece of data corresponding to "events of interest" is analyzed in depth. These events of interest are all the more complicated to abalyse that the signal to noise ratio (SNR) of M/EEG is poor.

1.1 Evoked potentials

In most traditional MEG or EEG cognitive experiments (notable exceptions are some events related to epilepsy which have a high SNR), stimuli are presented multiple times (each repetition is called a trial). The resulting measurements are aligned and averaged in order to improve the signal to noise ratio (evoked potentials).

Such a procedure assumes that signals do not vary across trials, and that they can easily be aligned, usually with respect to some "reference event". This reference event can be the onset of the stimulus presentation, or a measured subject reaction time. Such alignment is not always possible because latencies of the brain responses may vary across trials (and cumulative latencies make events far from the reference event more difficult to align). Averaging also erases many details in the signal, to the point that some components may disappear. Obviously, this may happen for events that are "not" time-locked to the reference event. Attention and habituation are other sources of variability across trials: while these may not be sufficient to make the events disappear in the average data, they can affect the strength of the signal or its perceived duration. More severe is the case of high frequency events that are not phase-locked across trials. Such events, even time-locked to a reference event, tend to cancel out in the average data and thus are difficult to

detect. They are usually best detected by averaging the time-power images corresponding to the signal introduced in [9].

1.2 Single trial analysis

Since the seminal work of Lehmann&al on microstates [7], much effort is being devoted in the community in order to be able to analyze single-trial measurements, or to segment continuous strands of data into pieces within which the signals enjoy similar properties. Statistical methods must be adapted to the multidimensionality of the data and heterogeneity of the dimensions (time, 3D space, trials, conditions, subjects). This analysis of single-trial measurements is of particular importance for Brain Computer interfaces, which have to detect such brain states in real time to transform them into commands.

1.3 Statistical tools used in M/EEG analysis

In this presentation, we will present some of these characteristics of EEG/MEG signals and review a certain number of techniques that have been used to analyse such datasets. While quite a few advanced techniques such as dictionary learning, deep-learning, optimal transport... have been explored recently [4, 3, 6, 5, 8, 2], it is remarkable how well simpler strategies such as LDA or CSP remain effective in practise, showing that some work is still needed to understand and efficiently exploit the actual content of such data. An remarkable exception is the Riemanian potatoe approach [1], which seems to consistently improve over the current state-of-the-art.

Bibliographie

References

[1] A. Barachant, A. Andreev, and M. Congedo. The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. In *TOBI Workshop IV*, pages 19–20, Sion, Switzerland, Jan. 2013.

[2] B. Belaoucha and T. Papadopoulo. Structural connectivity to reconstruct brain activation and effective connectivity between brain regions. *Journal of Neural Engineering*, 17(3):035006, June 2020.

[3] C. Bénar, M. Clerc, and T. Papadopoulo. Adaptive time-frequency models for single-trial M/EEG analysis. In Karssemeijer and Lelieveldt, editors, *Information Processing in Medical Imaging*, volume 4584 of *Lecture Notes in Computer Science*, pages 458–469. Springer, 2007.

[4] K. J. Blinowska, R. Kuś, and M. Kamiński. Granger causality and information flow in multivariate processes. *Phys. Rev. E*, 70(5):050902, 2004.

[5] N. Gayraud. *Adaptive machine learning methods for event related potential-based brain computer interfaces*. Theses, Université Côte d'Azur, Dec. 2018.

[6] S. Hitziger, M. Clerc, S. Saillet, C. Bénar, and T. Papadopoulo. Adaptive waveform learning: A framework for modeling variability in neurophysiological signals. *IEEE Transactions on Signal Processing*, 65(16):4324–4338, Apr. 2017.

[7] D. Lehmann and W. Skrandies. Spatial analysis of evoked potentials in man - a review. *Progr Neurobiol*, 23(3):227–250, 1984.

[8] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019.

[9] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier. Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *J. Neurosci.*, 16(13):4240–4249, 1996.

Preserved central model for faster bidirectional compression in distributed settings

Constantin Philippenko[1] & Aymeric Dieuleveut[2]

CMAP, École Polytechnique, Institut Polytechnique de Paris
[1] *constantin.philippenko@polytechnique.edu*
[2] *aymeric.dieuleveut@polytechnique.edu*

Résumé. L'objectif de cette communication est de décrire une nouvelle approche pour résoudre les contraintes de communication dans un problème d'apprentissage distribué avec un serveur central. Nous proposons et analysons un nouvel algorithme qui effectue une compression *bi-directionnelle* et atteint le même taux de convergence que les algorithmes utilisant uniquement la compression en liaison montante (i.e., des *travailleurs* locaux vers le serveur central). Pour obtenir cette amélioration, nous introduisons `MCM`, un algorithme tel que la compression sur la liaison descendante n'impacte *que les modèles locaux*, tandis que le modèle global est préservé. En conséquence, et contrairement aux travaux précédents, les gradients sur les serveurs locaux sont calculés sur des *modèles perturbés*. Par conséquent, les preuves de convergence sont plus difficiles à établir et nécessitent un contrôle précis de cette perturbation. Pour l'assurer, `MCM` combine en outre la compression des modèles avec un mécanisme de mémoire. Cette analyse ouvre de nouvelles portes, par exemple en incorporant des modèles randomisés dépendants des travailleurs et une participation partielle.

Mots-clés. Apprentissage fédéré, contraintes de communication, apprentissage distribué

Abstract. The goal of this communication is to describe a new approach to tackle communication constraints in a distributed learning problem with a central server. We propose and analyze a new algorithm that performs bidirectional compression and achieves the same convergence rate as algorithms using only uplink (from the local workers to the central server) compression. To obtain this improvement, we design `MCM`, an algorithm such that the downlink compression *only impacts local models*, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on *perturbed models*. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, `MCM` additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.

Keywords. Federated Learning, Communication Constraints, Distributed Learning

1 MCM, a new algorithm for bidirectional compression

In this work, we consider a setting using a central server that aggregates updates from remote nodes. Formally, we have a number of features $d \in \mathbb{N}^*$, and a convex cost function $F : \mathbb{R}^d \to \mathbb{R}$. We want to solve the following distributed convex optimization problem using stochastic gradient algorithms:

$$\min_{w \in \mathbb{R}^d} F(w) \text{ with } F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w) \,, \tag{1}$$

where $(F_i)_{i=1}^N$ is a *local* risk function (empirical risk or expected risk in a streaming framework). This applies to both instances of *distributed* and *federated* learning.

In the convex case, we assume there exists an optimal parameter w_*, and denote $F_* = F(w_*)$. To solve this problem, we rely on a stochastic gradient descent (SGD) algorithm. A stochastic gradient g_{k+1}^i is provided at iteration k in $\mathbb{N}$ to the device i in $[\![1, N]\!]$, with a bounded variance $\mathbb{E}[||g_{k+1}^i(w) - \nabla F_i(w)||^2] \leq \sigma^2$. This gradient oracle can be computed on a mini-batch of size b to improve parallelization. This function is then evaluated at point w_k. In the classical centralized framework (without compression), for a learning rate γ, SGD corresponds to:

$$w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} g_{k+1}^i(w_k) \,. \tag{2}$$

An important issue of those frameworks is the high communication cost between the workers and the central server [4, Sec. 3.5]. This cost is a concern from several points of view. First, exchanging information can be the bottleneck in terms of speed. Second, the data consumption and the bandwidth usage of training large distributed models can be problematic; and furthermore, the energetic and environmental impact of those exchanges is a growing concern. Over the last few years, new algorithms were introduced, compressing messages in the *upload communications* (i.e., from remote devices to the central server) in order to reduce the size of those exchanges More recently, a new trend has emerged to also compress the *downlink communication*: this is *bidirectional compression*, as illustrated in Figure 1.

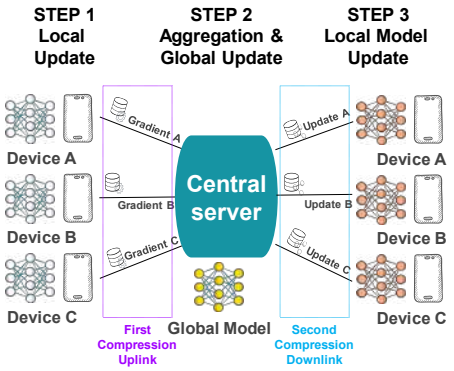


Figure 1: Illustration of the bidirectional compression framework with "model compression". At step 1, we send a gradient, at step 2 we communicate a model difference.

To perform downlink communication, existing bidirectional algorithms [11, 14, 9, 5, 8, 3, 13] first aggregate all the information they have received, compress them and then carry out the broadcast. Both the main "global" model and the "local" ones perform the *same* update

with this compressed information. Consequently, the model hold on the central server and the one used on the local workers (to query the gradient oracle) are identical. However, this means that the model on the central server has been artificially *degraded*: instead of using all the information it has received, it is updated with the compressed information.

Here, we focus on *preserving* (instead of *degrading*) the central model: the update made on its side does not depend on the downlink compression. This implies that the local models are *different* from the central model. The local gradients are thus measured on a *"perturbed model"* (or *"perturbed iterate"*: such an approach requires a more involved analysis and the algorithm must be carefully designed to control the deviation between the local and global models [6]. For example, algorithms directly compressing the model or the update would simply not converge.

We propose `MCM` - *Model Compression with Memory* - a new algorithm that 1) preserves the central model, and 2) uses a memory scheme to reduce the variance of the local model. We prove that the convergence of this method is similar to the one of algorithms using only unidirectional compression.

Bidirectional compression consists in compressing communications in both directions between the central server and remote devices. We use two potentially different compression operators, respectively $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ to compress the message in each direction, as illustrated in Figure 1. Roughly speaking, the update in eq. (2) becomes:

$$w_{k+1} = w_k - \gamma \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g_{k+1}^i(w_k))\right). \tag{3}$$

However, this approach has a major drawback. The central server receives and aggregates information $\frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g_{k+1}^i(w_k))$. But in order to be able to broadcast it back, it compresses it, *before* applying the update. We refer to this strategy as the "degraded update" approach. Its major advantage is simplicity, and it was used in all previous papers performing double compression. Yet, it appears to be a waste of valuable information. In this paper, we update the global model w_{k+1} independently of the downlink compression:

$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{up}}\left(g_{k+1}^i(\hat{w}_k)\right) \\ \hat{w}_{k+1} = C_{\text{dwn}}(w_{k+1}) \end{cases}. \tag{4}$$

However, bluntly compressing w_{k+1} in eq. (4) hinders convergence, thus the second part of the update needs to be refined by adding a memory mechanism. **We now describe both communication stages of the real `MCM`, as the combination of eqs. (5) and (6).**

Downlink Communication. We introduce a *downlink memory term* $(H_k)_k$, which is available on both workers and central server. The difference Ω_{k+1} between the model and this memory is compressed and exchanged, then the local model is reconstructed from

this information. The memory is then updated, with a learning rate α:

$$\left\{\begin{array}{l} \Omega_{k+1} = w_{k+1} - H_k\,, \\ \widehat{w}_{k+1} = H_k + \mathcal{C}_{\rm dwn}(\Omega_{k+1}) \\ H_{k+1} = H_k + \alpha \mathcal{C}_{\rm dwn}(\Omega_{k+1}). \end{array}\right. \tag{5}$$

Introducing this memory mechanism is crucial to control the variance of the local model $\widehat{w}_{k+1}$. To the best of our knowledge `MCM` is the first algorithm that uses such a memory mechanism for downlink compression. This mechanism was introduced by Mishchenko et al. [7] for the uplink compression.

Uplink Communication. The motivation to introduce an uplink memory term h_k^i for each device $i \in [\![1, N]\!]$ is different, and better understood. Indeed, for the uplink direction, this mechanism is only necessary (and then crucial) to handle heterogeneous workers [i.e., with different data distributions, see e.g. 8]. Here, the difference Δ_k^i between the stochastic gradient g_{k+1}^i at the local model $\widehat{w}_k$ (as defined in eq. (5)) and the memory term is compressed and exchanged. The memory is then updated with a rate β:

$$\left\{\begin{array}{l} \forall i \in [\![1, N]\!], \Delta_k^i = g_{k+1}^i(\widehat{w}_k) - h_k^i \\ w_{k+1} = w_k - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}_{\rm up}(\Delta_k^i) + h_k^i \\ h_{k+1}^i = h_k^i + \beta \mathcal{C}_{\rm up}(\Delta_k^i). \end{array}\right. \tag{6}$$

Remark 1 (Rate α)**.** *It is necessary to use $\alpha < 1$. Otherwise, the compression noise tends to propagate and is amplified, because of the multiplicative nature of the compression noise.*

As downlink communication can be more efficient than uplink, we consider distinct operators $\mathcal{C}_{\rm dwn}$, $\mathcal{C}_{\rm up}$ and allow the corresponding compressions levels to be distinct: there exist constants $\omega_{\mathcal{C}}^{\rm up}, \omega_{\mathcal{C}}^{\rm dwn} \in \mathbb{R}_+^*$, such that the compression operators $\mathcal{C}_{\rm up}$ and $\mathcal{C}_{\rm dwn}$ satisfy the two following properties for all w in $\mathbb{R}^d$:

$$\left\{\begin{array}{l} \mathbb{E}[\mathcal{C}_{\rm up/dwn}(w)] = w\,, \\ \mathbb{E}[\left\|\mathcal{C}_{\rm up/dwn}(w) - w\right\|^2] \leq \omega_{\mathcal{C}}^{\rm up/dwn} \left\|w\right\|^2\,. \end{array}\right.$$

The higher is $\omega_{\mathcal{C}}$, the more aggressive the compression is. This class of unbiased operators encompasses *sparsification*, *quantization* and *sketching*.

Remark 2 (Related work on Perturbed iterate analysis)**.** *The theory of perturbed iterate analysis was introduced by Mania et al. [6] to deal with asynchronous SGD. More recently, it was used by Stich and Karimireddy [10], Gorbunov et al. [2] to analyze the convergence of algorithms with uplink compressions, error feedback and asynchrony.*

Remark 3 (Related work on double compression)**.** *We summarize in Table 1 the main algorithms developed in order to carry out compression in distributed training and the novelties of our approach.*

Table 1: Features of the main existing algorithms performing compression. e_k^i (resp. E_k) denotes the use of error-feedback at uplink (resp. downlink). h_k^i (resp. H_k) denotes the use of a memory at uplink (resp. downlink). Note that `Dist-EF-SGD` is identical to `Double-Squeeze` but has been developed simultaneously and independently.

	Compr.	e_k^i	h_k^i	E_k	H_k	Rand.	update point
`Qsgd` Alistarh et al. [1]	one-way						
`ECQ-sgd` Wu et al. [12]	one-way	✓					
`Diana` Mishchenko et al. [7]	one-way		✓				
`Dore` Liu et al. [5]	two-way		✓	✓			degraded
`Double-Squeeze` Tang et al. [11]	two-way	✓		✓			degraded
`Dist-EF-SGD` Zheng et al. [14]	two-way	✓		✓			degraded
`Artemis` Philippenko and Dieuleveut [8]	two-way		✓				degraded
`MCM` & `Rand-MCM`	two-way		✓		✓	(✓)	non-degraded

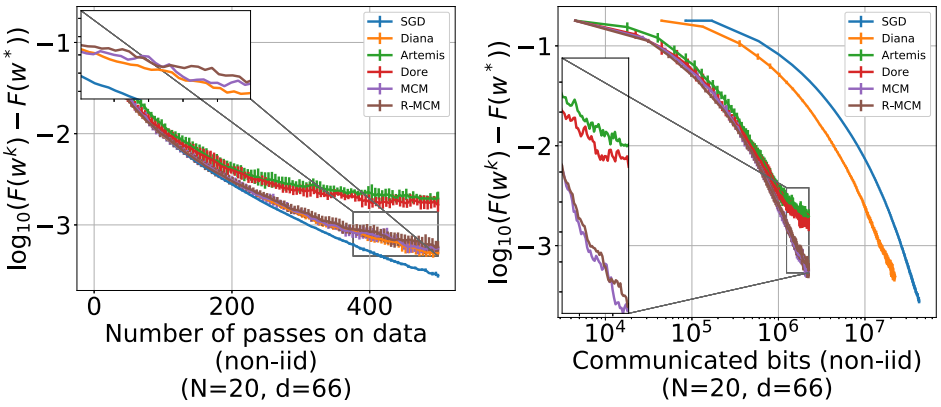
Remark 4 (The randomization mechanism, `Rand-MCM`)**.** *We can extend the framework to* ***`Rand-MCM`****, that includes randomization. It consists in performing an independent compression for each device instead of performing a single one for all of them. As a consequence, each worker holds a different model centered around the global one. This introduces some supplementary randomness that stabilizes the algorithm. Formally, we will consider N mutually independent compression operators $\mathcal{C}_{\mathrm{dwn},i}$ instead of a single one $\mathcal{C}_{\mathrm{dwn}}$, and the central server will send to the device i at iteration $k+1$ the compression of the difference between its model and the local memory on worker i: $\mathcal{C}_{\mathrm{dwn},i}(w_{k+1} - H_k^i)$. This allows to incorporate worker dependent compression or partial participation.*

2 Convergence results for `MCM`

We can provide convergence guarantees for `MCM` in three regimes: convex, strongly-convex and non-convex. The analysis relies on controlling the variance of the local model with respect to the global model depending on all previous gradients. The memory term here plays a key role: without memory, the variance would increase with the number of iterations. On the other hand, the learning rate α for the memory term has to be controlled carefully. Indeed, if α is too large, the variance diverges: this phenomenon is similar to the divergence observed in frameworks involving error feedback, when the compression operator is not contractive.

Theorem 1. *Consider the* ***`MCM`*** *update as in eq. (4). For an L-smooth function F, if $\gamma \leq 1/(8\omega_{\mathcal{C}}^{\mathrm{dwn}} L)$ and $\alpha \leq 1/(4\omega_{\mathcal{C}}^{\mathrm{dwn}})$, then for all k in $\mathbb{N}$:*

$$\mathbb{E}[\|w_k - \widehat{w}_k\|^2] \leq \frac{4\omega_{\mathcal{C}}^{\mathrm{dwn}}\gamma^2\sigma^2(1+\omega_{\mathcal{C}}^{\mathrm{up}})}{Nb\alpha} + 2\omega_{\mathcal{C}}^{\mathrm{dwn}}\gamma^2\left(\frac{1}{\alpha} + \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha}{2}\right)^{k-t}\mathbb{E}\left\|\nabla F(\widehat{w}_{t-1})\right\|^2.$$



(a) X-axis in # iterations. (b) X-axis in # bits.

Figure 2: Quantum with $b = 400$, $\gamma = 1/L$. Best seen in colors.

The proof requires to control the different terms that contribute in the local model's variance. We can then derive the following convergence rate in the convex case.

Theorem 2 (Convergence of `MCM` in the convex case)**.** *For an L-smooth function F, for a given K in $\mathbb{N}$ large enough, a step size $\gamma = 1/(L\sqrt{K})$, a given learning rate $\alpha = 1/(8\omega_{\mathcal{C}}^{\mathrm{dwn}})$, after running K iterations, we have, with $\Phi = (1 + \omega_{\mathcal{C}}^{\mathrm{up}})\left(1 + \frac{64(\omega_{\mathcal{C}}^{\mathrm{dwn}})^2}{\sqrt{K}}\right)$:*

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq \frac{\|w_0 - w_*\|^2 L}{\sqrt{K}} + \frac{\sigma^2 \Phi}{NbL\sqrt{K}}\,.$$

Observe that the downlink compression constant only appears in the second-order term, scaling as $1/K$. In other words, the convergence rate is equivalent to the convergence rate of `Diana`. It is also possible to provide convergence guarantees in the strongly-convex case with steps scaling as $((\mu k)^{-1})_{1 \leq k \leq K}$), and non-convex case.

3 Experiments

We here illustrate the performance of `MCM`on *quantum* a real dataset used for logistic regression in a non-i.i.d. settings: 50.000 points, $d = 66$ features, $N = 20$ devices, $\gamma = 1/L$. There are between 900 and 10500 points by devices, with a median at 2300. We compare `MCM` with classical algorithms used in distributed settings: `Diana`, `Artemis`, `Dore` and of course the simplest setting - `SGD`, which is the baseline. We plot the log of the excess loss $F(w_k) - F_*$, averaged on 5 runs, with error bars displayed on each figure (but not in the "zoom square"), corresponding to the standard deviation of $\log_{10}(F(w_k) - F_*)$.

References

[1] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30: 1709–1720, 2017.

[2] E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtárik. Linearly Converging Error Compensated SGD. *arXiv:2010.12292 [cs, math]*, October 2020. arXiv: 2010.12292.

[3] S. Horváth and P. Richtárik. A Better Alternative to Error Feedback for Communication-Efficient Distributed Learning. *arXiv:2006.11077 [cs, stat]*, June 2020. arXiv: 2006.11077.

[4] P. e. a. Kairouz. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. arXiv: 1912.04977.

[5] X. Liu, Y. Li, J. Tang, and M. Yan. A Double Residual Compression Algorithm for Efficient Distributed Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143, June 2020. ISSN: 1938-7228 Section: Machine Learning.

[6] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. 07 2015. doi: 10.1137/16M1057000.

[7] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed Learning with Compressed Gradient Differences. *arXiv:1901.09269 [cs, math, stat]*, June 2019. arXiv: 1901.09269.

[8] C. Philippenko and A. Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in Federated Learning. *arXiv:2006.14591 [cs, stat]*, November 2020. arXiv: 2006.14591.

[9] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2019. ISSN 2162-2388. doi: 10.1109/TNNLS.2019.2944481. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[10] S. U. Stich and S. P. Karimireddy. The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Communication. *arXiv:1909.05350 [cs, math, stat]*, September 2019. arXiv: 1909.05350.

[11] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, May 2019. ISSN: 2640-3498.

[12] J. Wu, W. Huang, J. Huang, and T. Zhang. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In *International Conference on Machine Learning*, pages 5325–5333. PMLR, July 2018. ISSN: 2640-3498.

[13] A. Xu, Z. Huo, and H. Huang. Training Faster with Compressed Gradient. *arXiv:2008.05823 [cs, stat]*, August 2020. arXiv: 2008.05823.

[14] S. Zheng, Z. Huang, and J. Kwok. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11450–11460. Curran Associates, Inc., 2019.

Categorical functional data analysis with the cfda R package

Cristian Preda [1], Quentin Grimonprez [2], Vincent Vandewalle [3]

[1] *Univ. Lille, CNRS, UMR 8524, Inria - Laboratoire Paul Painlevé, F-59000 Lille, France. cristian.preda@univ-lille.fr*

[2] *DiagRAMS Technologies, Inria. qgrimonprez@diagrams-technologies.com*

[3] *Univ. Lille, CHU Lille, ULR 2694 - METRICS : Évaluation des technologies de santé et des pratiques médicales, Inria, F-59000 Lille, France. vincent.vandewalle@univ-lille.fr*

Résumé. Nous présentons comment prendre en compte des données fonctionnelles qualitatives, représentées par les trajectoires d'un processus de saut avec un temps continu et un ensemble fini d'états. En tant qu'extension de l'analyse des correspondances multiples à un ensemble infini de variables, les codages optimaux des états dans le temps sont approchés sur une base arbitraire finie de fonctions. Cela permet de réduire la dimension, d'optimiser la représentation et de visualiser des données dans les espaces dimension plus petites. Nous avons implémenté cette méthodologie dans le package R **cfda** disponible sur le CRAN, nous présenterons dans cette communication comment celle-ci peut être mise en œuvre sur des données réelles dans un contexte de classification non supervisée.

Mots-clés. données fonctionnelles ; données qualitatives ; processus stochastique ; analyse des correspondances multiples.

Abstract. We present how to take into account categorical functional data, represented by the trajectories of a jump process with a continuous time and a finite set of states. As an extension of multiple correspondence analysis to an infinite set of variables, the optimal codings of the states over time are approximated on an arbitrary finite function basis. This allows dimension reduction, representation optimization and visualisation of data in smaller dimensional spaces. We have implemented this methodology in the R package **cfda** available on the CRAN, we will present in this communication how it can be implemented on real data in an clustering framework.

Keywords. functional data; categorical data; stochastic process; multiple correspondence analysis.

1 Introduction to categorical functional data

Most literature devoted to functional data considers data as sample paths of a real-valued stochastic process, $X = \{X_t, t \in \mathcal{T}\}$, $X_t \in \mathbb{R}^p$, $p \geq 1$ where $\mathcal{T}$ is some continuous set. Among a considerable record of papers on the subject, the monographs of Ramsay and

Silverman [3] among others still remains a reference presenting the main methodologies for visualisation, denoising, classification and regression when dealing with functional data represented by real-valued functions.

We consider the case where the underlying stochastic model generating the data is a continuous-time stochastic process $X = \{X_t, t \in \mathcal{T}\}$ such that for all $t \in \mathcal{T}$, X_t is a categorical random variable rather than a real-valued one.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{S} = \{s_1, \ldots, s_m\}$, $m \geq 2$, be a set of m states and $X = \{X_t \ ; \ X_t : \Omega \longrightarrow \mathcal{S}, \quad t \in \mathcal{T}\}$ be a family of categorical random variables indexed by $\mathcal{T}$. Thus, a path of X is a sequence of states s_{i_j} and times points t_i of transitions from one state to another one : $\{(s_{i_1}, t_1), (s_{i_2}, t_2), \ldots\}$, with $s_{i_j} \in \mathcal{S}$ and $t_i \in \mathcal{T}$.

We call the sample paths of the process *categorical functional data*. Such type of data is able to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on.

In Section 2 we present the theoretical background of the optimal encoding methodology defining the *principal components* of the process X throughout the optimal encodings. The approximation of the optimal encodings of the states into a basis of functions and optimal representation of categorical functional data in lower dimensional spaces are detailed. The implementation of the optimal encodings is presented throughout the **cfda** R package in Section 3 where an application on a real data set (care trajectories for patients diagnosed with severe infection) is performed in view of visualisation and clustering.

2 Extension of multiple correspondence analysis

Optimal encoding Without loss of generality, let suppose that $\mathcal{T} = [0, T]$, with $T > 0$. For $x, y \in \mathcal{S}$ and $\forall t \in [0, T]$, let denote by:

- $\mathbf{1}_t^x = 1$ if $X_t = x$, and 0 otherwise,
- $p^x(t) = \mathsf{P}(X_t = x)$ and $p^{x,y}(t, s) = \mathsf{P}(X_t = x, X_s = y)$.

The general hypotheses considered in that framework are:

H_1**:** the process X is continuous in probability, $\lim_{h \to 0} \mathsf{P}(X_{t+h} \neq X_t) = 0$

H_2**:** for each time $t \in [0, T]$ (except possibly a finite discrete set of time points), any state has a strictly positive probability to occur: $p^x(t) \neq 0, \forall x \in \mathcal{S}, \forall t \in [0, T]$.

In this framework, Deville [1] extends the multiple correspondence analysis to the process X (seen as infinite random variables). This is related the following eigen-value problem

$$\int_0^T \sum_{y \in \mathcal{S}} p^{x,y}(t, s) a^y(s) ds = \lambda a^x(t) p^x(t), \quad \forall t \in [0, T], \forall x \in \mathcal{S}, \tag{1}$$

where $\{a^x\}_{x\in\mathcal{S}}$ are deterministic functions on $[0,T]$ that we call *optimal encoding* functions. Under the hypothesis H_1 and H_2 it admits the sequence of eigenvalues $\{\lambda_j\}_{j\geq 1}$ associated to the optimal encoding eigen-functions $\{a_j^x, x\in\mathcal{S}\}_{j\geq 1}$. The j-th principal component z_j is derived from the j-th optimal encoding functions $\{a_j^x\}$ as

$$z_j = \int_0^T \sum_{x\in\mathcal{S}} a_j^x(t)\mathbf{1}_t^x dt, \quad \forall i \geq 1. \tag{2}$$

Dimension Reduction. Let $q \geq 1$, one obtains the best approximation of order q of X (viewed as a vector process $X = \{\mathbf{1}^x, x\in\mathcal{S}\}$) under the L_2 norm, among all the linear expansions of type

$$\mathbf{1}_t^x \approx \sum_{j=1}^{q} z_j a_j^x(t)\frac{1}{p^x(t)}, \quad \forall x\in\mathcal{S}.$$

Thus, the q first principal components,

$$\{z_1,\ldots,z_q\}, \quad q\geq 1,$$

allow for

- graphical representation of sample paths of X in $\mathbb{R}^q$ (especially for $q=2$, one obtains a 2-D representation of categorical functional data),
- fit of clustering and regression models with X as explanatory variables.

Discussion Technical details are not given here but can be found in [2]. The main idea is to consider and expansion of the $\{a^x\}_{x\in\mathcal{S}}$ on some basis, limiting the problem to some finite dimension problem thus solving a classical eigen-problem. One major interest for such dimension reduction is that it permits to consider data in $\mathbb{R}^q$ rather than the initial functional categorical data, and it gives a solution to perform clustering.

3 The cfda Package through an example

The **cfda** R package (available on the CRAN) provides functions to analyze categorical functional data allowing to compute basic statistics such as transition tables or visualisation, and compute the optimal encodings. It uses **ggplot2** package to display graphics and the **parallel** and **pbapply** packages for code parallelization. Other packages for analyzing sequences of categorical data exist, but not in a functional way.

Real Dataset The **cfda** package is illustrated with the `care` dataset. It contains 2929 care trajectories for patients diagnosed with a severe infection. Each month from the diagnosis of the infection, the follow-up of each patient is filled in using one of the following 4 states: "`D`", the patient has not a medical follow-up, "`C`", the patient has a medical follow-up but no treatment, "`T`", the patient has a medical follow-up with a treatment but the infection is not suppressed and "`S`", the patient has a medical follow-up with a treatment and the infection is suppressed. A sample of the individuals from the `care` dataset is plotted using the `plotData` function (see Figure 1a). Each line corresponds to an individual of the dataset, the successive changes of states are represented by different colors.

```
R> data(care)
```

Extract a Dataset Meeting the Constraints To compute the encodings, each individual must have the same start and end time. This is not the case in the `care` dataset. So, we select patients with a follow-up of at least 18 months and works from $t = 0$ to $t = 18$ months. To restrict individuals to a maximal time value of 18 months, we use the `cut_data` function that has two parameters: `data` and `Tmax`, the maximal time value. After applying this function, all individuals have `Tmax` as ending time.

```
R> duration <- compute_duration(care)
R> idToKeep <- names(duration[duration >= 18])
R> care2 <- cut_data(care[care$id %in% idToKeep, ], 18)
```

Optimal Encoding While it permits to performs basic statistics and visualisations, the main contribution of **cfda** is the computation of an optimal encoding for categorical functional data performed by the `compute_optimal_encoding` function. The two main parameters are `data`, the dataset in the **cfda** format, and `basisobj`, a `basisfd` object created using the different `create.*.basis` functions from the **fda** package. It also performs bootstrap for computing a confidence interval on the computed encoding; associated parameters are `computeCI`, a logical indicating if bootstrap must be performed, `nBootstrap`, the number of bootstrap samples, and `propBootstrap`, the proportion of individuals used for each bootstrap sample. Other parameters are `nCores` the number of cores to use, `verbose`, if `TRUE`, some information are printed during the process.

```
R> set.seed(42)
R> basis <- create.bspline.basis(c(0, 18), nbasis = 10, norder = 4)
R> fmca <- compute_optimal_encoding(care2, basis, computeCI = TRUE, nCores = 7)
```

Plot Functions Three plot functions are associated with the `compute_optimal_encoding` function, the first argument of these functions is the output of `compute_optimal_encoding`.

The first one, the `plot` function plots the encodings associated with a given eigenvector (`harm` parameter, by default, the encodings associated with the first eigenvector are plotted). If `compute_optimal_encoding` was run with parameter `computeCI = TRUE`, then the confidence interval can be added on the plot using the parameter `addCI = TRUE`. A subset of the states can be plotted by providing a vector with the state names to the `states` parameter. The `plotEigenvalues` function plots the computed eigenvalues. The last one is the `plotComponent` function that plots the individuals using the given components (`comp` parameter, a vector of length 2 containing the components' number). The encodings are displayed in Figure 1c and the plot of the first factorial map is presented in Figure 1b.

```
R> plotComponent(fmca, comp = c(1, 2), addNames = FALSE)
R> plot(fmca, addCI = TRUE)
```

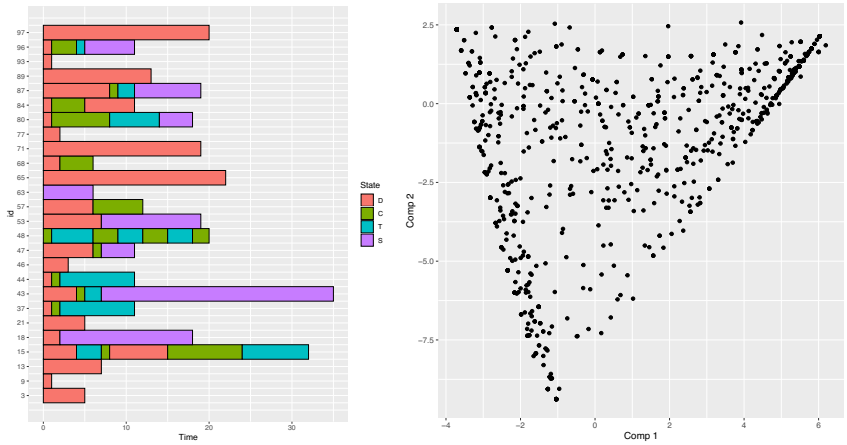
Application to Clustering The proposed method produces numerical encoding for categorical functional data. These encoding can be used for classical statistical methods such that regression or clustering. In the following, we perform a hierarchical clustering to find a structure in the `care` dataset. The clustering is performed with the first principal components explaining at least 90% of the variance. The different clusters are associated with the time spent in the different states after leaving the state `"D"` (cf. Figure 2). For example, the cluster number 1 corresponds to individuals that have spent most of their time (after `"D"`) in the `"C"` state.

References

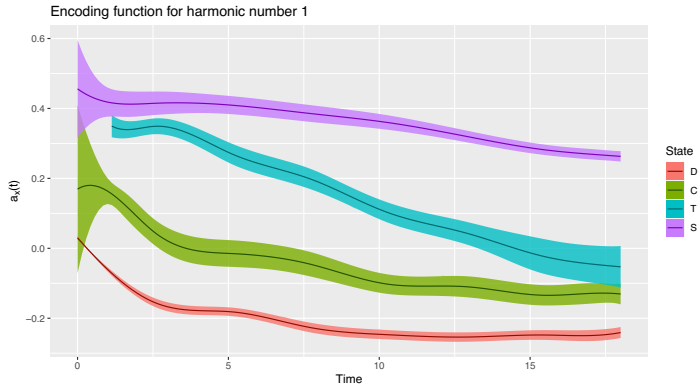
[1] Deville, J.-C. (1982). Analyse de données chronologiques qualitatives : Comment analyser des calendriers ? *Annales de l'INSEE*, 45:45–104.

[2] Preda, C., Grimonprez, Q., and Vandewalle, V. (2020). cfda: an R Package for Categorical Functional Data Analysis. working paper or preprint `https://hal.inria.fr/hal-02973094`.

[3] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer-Verlag New York.



(a) `Sample of individuals of the care dataset` (b) `plotComponent`



(c) `plot(fmca, addCI = TRUE)`

Figure 1 – Plots generated of sample of data and of the first factorial map.

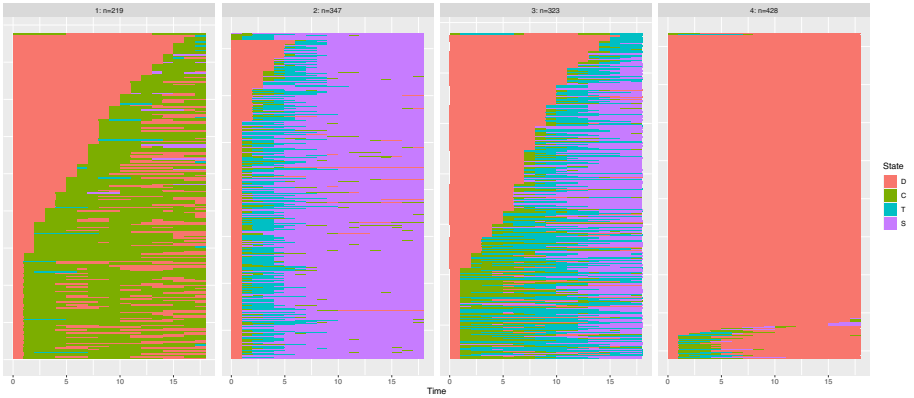


Figure 2 – Content of the different clusters.

Titre : Justice et IA, un dialogue à éclaircir

Auteur : Laurence Pécaut-Rivolier
Conseiller à la chambre sociale, Cour de cassation

Résumé :
La question de l'intelligence artificielle se pose de manière tout à fait spécifique s'agissant du processus juridictionnel. La première particularité est que l'institution judiciaire a été, pour diverses raisons, très longtemps réfractaire à toute introduction de l'intelligence artificielle si bien qu'elle en méconnait fortement le fonctionnement et ne s'est pas, jusqu'à récemment, impliquée dans les réflexions sur l'apport de l'intelligence artificielle en matière de justice. Ce sont donc les acteurs extérieurs, cabinets d'avocats, instituts privés, qui ont élaboré et proposé des approches des décisions judiciaires par l'IA, en mettant en avant des objectifs qui ne sont pas toujours ou pas exclusivement ceux poursuivis par l'institution judiciaire. La seconde particularité vient du processus décisionnel lui-même. Bâti sur le syllogisme judiciaire, il nécessite pour un grand nombre d'affaires un mode de raisonnement et une adaptation à chaque cas d'espèce qui n'est pas toujours compatible avec l'IA . Aujourd'hui, alors que les acteurs judiciaires commencent à s'intéresser à l'IA, la question est donc de bien déterminer ce qui relève de l'apport de ce qui relève du risque, et d'avoir enfin une approche pluridisciplinaire en ce domaine.

L'AdaptSgenoLasso, une variante du SgenoLasso, pour la localisation de gènes et la prédiction génomique à l'aide des extrêmes

Charles-Elie Rabier [1,2] & Céline Delmas [3]

[1] *ISEM, Université de Montpellier, CNRS, EPHE, IRD, France*
[2] *IMAG, Université de Montpellier, CNRS, France*
charles-elie.rabier@umontpellier.fr
[3] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France*
celine.delmas.toulouse@inrae.fr

Résumé. Nous présentons l'AdaptSgenoLasso, une nouvelle méthode de vraisemblance pénalisée pour la localisation de gènes, et qui s'avère être une variante du SgenoLasso. L'AdaptSgenoLasso repose sur le concept d'un génotypage sélectif qui est autorisé à varier le long du génome. La version classique du génotypage sélectif, sur laquelle est basé le SgenoLasso, consiste à génotyper uniquement les individus extrêmes, afin d'augmenter le signal lié aux gènes. Cependant, comme le même pourcentage de sélection est appliqué à chaque position du génome, le signal se voit augmenté d'un même facteur proportionnel sur l'ensemble du génome. En considérant un génotypage sélectif qui varie le long du génome grâce à l'AdaptSgenoLasso, nous permettons aux généticiens d'imposer plus de poids à certains loci d'intérêt, connus pour être responsables de la variation de caractères quantitatifs. Le signal est désormais propre à chaque locus.

Mots-clés. Apprentissage statistique, Détection de gènes, Processus gaussien, Test d'hypothèses, Génotypage sélectif, Extrêmes

Abstract. We introduce here the AdaptSgenoLasso, a new penalized likelihood method for gene mapping, which is a modified version of the SgenoLasso. AdaptSgenoLasso relies on the concept of a selective genotyping that varies along the genome. The "original version" of the selective genotyping on which the SgenoLasso is built on, consists in genotyping only extreme individuals, in order to increase the signal from genes. However, since the same amount of selection is applied at all genome locations, the signal is increased of the same proportional factor everywhere. By considering a selective genotyping that varies along the genome thanks to the AdaptSgenoLasso, we allow geneticists to impose more weights on some loci of interest, known to be responsible for variation of the quantitative trait. The resulting signal is now dedicated to each locus.

Keywords. Statistical learning, Gene detection, Gaussian process, Hypothesis testing, Selective genotyping, Extremes

1 Contexte

On étudie une population backcross ($A \times B$) où A et B sont deux lignées homozygotes pures. On considère le problème de la détection de loci codant pour un caractère quantitatif, aussi appelés QTL (Quantitative Trait Loci), sur un chromosome donné. Le caractère est observé sur n individus et on note Y_j, $j = 1, \cdots, n$, les observations que l'on suppose i.i.d. Le mécanisme de la méïose fait que parmi les deux chromosomes d'un individu, un est purement hérité de A alors que l'autre est formé de morceaux de A et de morceaux de B du fait des crossing-overs. Le chromosome est représenté par le segment $[0, T]$. La distance sur $[0, T]$ est appelée distance génétique et est mesurée en Morgans. Le génome $X(t)$ d'un individu prend la valeur +1 si le chromosome recombiné est originaire de A à la position t et prend la valeur -1 s'il est originaire de B. Le modèle admis pour la structure stochastique de $X(\cdot)$ est dû à Haldane (1919):

$$X(0) \sim \frac{1}{2}(\delta_{+1} + \delta_{-1}), \quad X(t) = X(0)(-1)^{N(t)}$$

où $N(\cdot)$ est le processus de Poisson standard sur $[0, T]$ représentant le nombre de crossing-overs. De plus on suppose que m QTLs additifs influent sur le caractère quantitatif Y. On note q_s et $t_s^\star$ l'effet et la position du $s$ième QTL, $s = 1 \cdots m$. On suppose un modèle d'analyse de variance pour Y:

$$Y = \mu + \sum_{s=1}^{m} X(t_s^\star) q_s + \sigma\varepsilon \tag{1}$$

où ε est un bruit blanc gaussien. Dans le problème classique de détection de QTLs, l'"information génome" est disponible uniquement à des positions fixes $t_1 = 0 < t_2 < ... < t_K = T$, appelées marqueurs génétiques. Ainsi, d'ordinaire, une observation est (Y, $X(t_1)$, ..., $X(t_K)$) et le challenge réside dans le fait que le nombre m de QTLs m et leurs positions $t_1^\star$, ..., $t_m^\star$ sont inconnues. On notera $t^\star = (t_1^\star, ..., t_m^\star)$. Dans cette étude, nous considérons le problème classique, mais afin de réduire les coûts dus au génotypage, un génotypage sélectif qui varie le long du génome est considéré. Décrivons tout d'abord le concept du génotypage sélectif dans sa version originale, celle qui ne varie pas le long du génome. Le génotypage sélectif consiste à génotyper (i.e. obtenir l'information génétique aux marqueurs $X(t_1), \cdots, X(t_K)$), uniquement les individus extrêmes (i.e. les individus dont le phénotype Y est au delà d'un certain seuil: $Y \notin [S_-^1, S_+^1]$). Ce dispositif proposé par Lebowitz et al. (1987) s'avère très employé en agronomie, car il permet d'optimiser le génotypage et d'améliorer la puissance de détection. Afin d'introduire le génotypage sélectif qui varie le long du génome, considérons désormais quatre seuils S_-^1, S_-^2, S_+^2, S_+^1 appartenant à $\mathbb{R}$ tels que $S_-^1 \leq S_-^2 \leq S_+^2 \leq S_+^1$. Comme dans le cadre du génotypage sélectif classique, nous observons l'information génome à tous les marqueurs si et seulement si Y est extrême, à savoir si $Y \leq S_-^1$ ou $Y \geq S_+^1$. Cependant, nous considérons

également une carte génétique discrète contenant seulement quelques marqueurs appartenant à la carte dense originale (i.e. la carte comprenant tous les marqueurs), et nous observons à ces quelques positions, l'information génome des individus pour lesquels $Y \leq S_-^2$ ou $Y \geq S_+^2$. En d'autres termes, à ce nombre restreint de marqueurs, nous recueillons l'information génome d'un plus grand nombre d'individus extrêmes. Intuitivement, cela permet d'imposer des poids plus importants à certains loci (cf. Section 2) correspondant à des gènes majeurs bien connus par les généticiens.

Afin de décrire les deux cartes génétiques plus précisement, notons $\mathbb{T}_K^1$ l'ensemble $\{t_1, \ldots, t_K\}$ de marqueurs sur la carte dense, et $\mathbb{T}_K^2$ un sous espace de $\mathbb{T}_K^1$ (i.e. $\mathbb{T}_K^2 \subseteq \mathbb{T}_K^1$), représentant les marqueurs sur la carte discrète. On notera $\#\mathbb{T}_K^2$ le cardinal de $\mathbb{T}_K^2$, et $\sigma(.)$ la fonction injective telle que $\sigma : \{1, \ldots, \#\mathbb{T}_K^2\} \to \{1, \ldots, K\}$. De plus, nous imposons $\sigma(k) < \sigma(k')$ pour $k < k'$. Ainsi, $\mathbb{T}_K^2$ désigne l'ensemble $\left\{t_{\sigma(1)}, t_{\sigma(2)}, \ldots, t_{\sigma(\#\mathbb{T}_K^2)}\right\}$. Nous imposerons également $\sigma(1) = 1$ et $\sigma(\#\mathbb{T}_K^2) = K$, de telle sorte que les marqueurs positionnés en 0 et en T (i.e. aux extrémités du chromosome) soient également localisés sur la carte discrète. Si nous notons $\overline{X}(t)$ et $\widetilde{X}(t)$ les variables aléatoires telles que $\overline{X}(t) = X(t)1_{Y \notin \left[S_-^1\ ,\ S_+^1\right]}$ et $\widetilde{X}(t) = X(t)1_{Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]}$, alors dans notre problème, une observation correspond à la donnée de

$$\left(Y,\ \overline{X}(t_1),\ \overline{X}(t_2),\ ...,\ \overline{X}(t_K),\ \widetilde{X}(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)})\right).$$

Avec nos notations,

- quand $Y \notin \left[S_-^1\ ,\ S_+^1\right]$, nous avons $\overline{X}(t_1) = X(t_1),\ ..., \overline{X}(t_K) = X(t_K)$, ce qui signifie que l'information génome est connue sur la carte dense $\mathbb{T}_K^1$ (et a fortiori sur la carte sparse $\mathbb{T}_K^2$).

- quand $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$, nous avons $\widetilde{X}(t_{\sigma(1)}) = X(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}) = X(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)}) = X(t_{\sigma(\#\mathbb{T}_K^2)})$, à savoir l'information génome est connue uniquement sur la carte sparse $\mathbb{T}_K^2$.

- quand $Y \in \left[S_-^2\ ,\ S_+^2\right]$, nous avons $\overline{X}(t_1) = 0,\ ..., \overline{X}(t_K) = 0$, et $\widetilde{X}(t_{\sigma(1)}) = 0, \widetilde{X}(t_{\sigma(2)}) = 0,\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)}) = 0$, ce qui signifie que l'information génome est manquante à tous les marqueurs (i.e. $\mathbb{T}_K^1$).

Nous observons dès lors n observations
$\left(Y_j,\ \overline{X}_j(t_1),\ \overline{X}_j(t_2),\ ...,\ \overline{X}_j(t_K),\ \widetilde{X}_j(t_{\sigma(1)}), \widetilde{X}_j(t_{\sigma(2)}),\ ..., \widetilde{X}_j(t_{\sigma(\#\mathbb{T}_K^2)})\right)$ pour $j = 1, \cdots, n$; supposées iid.

Avant de décrire nos résultats et notre nouvelle méthode de sélection de variables, rappelons le concept de l'Interval Mapping (Lander et Botstein, 1989). Lorsqu'un seul QTL est positionné sur le chromosome (i.e. $m = 1$ dans la formule (1)), l'Interval

Mapping consiste à calculer le Test du Rapport de Vraisemblance (LRT) à chaque position $t \in [0, T]$, confrontant l'hypothèse nulle d'absence de QTL H_0:"$q_1 = 0$," contre l'alternative "$q_1 \neq 0$,". Cela conduit à un processus de LRT $\Lambda_n(.)$ et à un processus de score $S_n(.)$. Ces processus ont été étudiés en détail dans le passé dans la situation de données complètes où tous les individus sont génotypés (e.g. Cierco C., 1998, Azaïs et Wschebor, 2009, Chang et al., 2009, Azaïs et al, 2012), et plus tard dans le cadre du génotypage sélectif classique (e.g Rabier, 2014, 2015). Le maximum de ces processus correspond au LRT sur l'ensemble du chromosome, et la distribution asymptotique de ces processus est désormais bien connue. Cependant la statistique du maximum s'avère inappropriée lorsque $m > 1$. Ainsi, dans cette étude, nous proposons d'étudier, dans le cadre du génotypage sélectif qui varie, la distribution asymptotique des processus de LRT et de score sous l'alternative générale de m QTLs sur le génome. Cela nous permettra d'introduire une nouvelle méthode de sélection de variables, l'AdaptSgenoLASSO, afin d'identifier les nombreux QTLs le long du génome, à l'instar de la méthode du SgenoLasso (Rabier et Delmas, 2021), proposée récemment dans le cadre du génotypage sélectif classique. Contrairement au Lasso (Tibshirani, 1996), le SgenoLasso présente l'avantage de gérer efficacement les données extrêmes. De plus, nous avons montré qu'il présentait de meilleurs performances que le récent RaLasso (Fan et al., 2017) qui modélise pourtant les dépendances entre les erreurs et les regresseurs. Ainsi, l'AdaptSgenoLasso, la nouvelle variante du SgenoLasso permettant d'accorder plus d'importance à certains loci bien connus, s'avère prometteuse.

Introduisons les notations suivantes:

$$\gamma_1 := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^1,\, S_+^1\right]\right) \ ,\ \gamma_1^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^1\right) \ ,\ \gamma_1^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^1\right) \ , \tag{2}$$

$$\gamma := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^2,\, S_+^2\right]\right) \ ,\ \gamma^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^2\right) \ ,\ \gamma^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^2\right) \ , \tag{3}$$

$$\mathcal{A}_1 := \sigma^2 \left\{ \gamma_1 + z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-}) \right\} \ , \tag{4}$$

$$\mathcal{B} := \sigma^2 \left\{ \gamma + z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-}) \right\} \ ,\ \mathcal{A}_2 := \mathcal{B} - \mathcal{A}_1 \tag{5}$$

où $\varphi(x)$ et z_α désignent respectivement la densité d'une loi normale standard prise au point x et le quantile d'ordre $1-\alpha$ d'une loi normale standard.
Nos principaux résultats sont résumés dans la section suivante.

2 Résultats

Dans ce qui suit, on considère des valeurs de t distinctes des positions de marqueurs, i.e. $t \in [t_1, t_K] \backslash \mathbb{T}_K^1$. Pour $i = 1, 2$, on définit $t^{\ell,i}$ and $t^{r,i}$ de la manière suivante:

$$t^{\ell,i} = \sup\left\{t_k \in \mathbb{T}_K^i : t_k < t\right\} \ \ ,\ \ t^{r,i} = \inf\left\{t_k \in \mathbb{T}_K^i : t < t_k\right\}. \tag{6}$$

En d'autres termes, en fonction de la carte, t appartient à l'intervalle de marqueurs $(t^{\ell,1}, t^{r,1})$ ou $(t^{\ell,2}, t^{r,2})$.

Theorem 1 *Supposons que les paramètres* $(q_1, ..., q_m, \mu, \sigma^2)$ *varient dans un compact, que* $\exists b > 0$ *tel que* $\sigma^2 \geq b > 0$*, et que* m *est fini. Soit* H_0 *l'hypothèse nulle d'absence de QTL sur* $[0, T]$ *et définissons les hypothèses alternatives suivantes:*

$$H_{a,t^\star} : \text{“il y a } m \text{ QTL localisés respectivement en } t_1^\star, \ ..., \ t_m^\star$$
$$\text{d'effets } q_1 = a_1/\sqrt{n}, ..., q_m = a_m/\sqrt{n} \text{ où } a_1 \neq 0, ..., a_m \neq 0\text{”} .$$

Alors lorsque n *tend vers l'infini, les processus* $S_n(.)$ *et* $\Lambda_n(.)$ *vérifient :*

$$S_n(.) \Rightarrow Z(.) \ , \ \Lambda_n(.) \overset{F.d.}{\to} Z^2(.) \ , \ \sup \Lambda_n(.) \xrightarrow{\mathcal{L}} \sup Z^2(.) \tag{7}$$

sous $\mathcal{H}_0$ *et* $\mathcal{H}_{a,t^\star}$*, où* $\Rightarrow$ *et F.d. désignent respectivement la convergence faible et la convergence des lois finies-dimensionnelles, et où* $Z(.)$ *est le processus Gaussien de variance* 1 *tel que* $\forall t \in [t_1, t_K] \backslash \mathbb{T}_K^1$ *:*

$$Z(t) = \frac{\sqrt{\mathcal{A}_1} \ \xi_1(t) V_1(t) \ + \ \sqrt{\mathcal{A}_2} \ \xi_2(t) V_2(t)}{\sqrt{\mathcal{A}_1 \ \xi_1^2(t) + \mathcal{A}_2 \ \xi_2^2(t)}} \ .$$

$V_1(.)$ *et* $V_2(.)$ *sont des processus Gaussiens indépendants de variance* 1 *tels que*

$$V_i(t) = \left\{ \alpha_i(t) \ V_i(t^{\ell,i}) \ + \ \beta_i(t) \ V_i(t^{r,i}) \right\} / \xi_i(t)$$
$$\forall (t_k, t_{k'}) \in \mathbb{T}_K^i \times \mathbb{T}_K^i \ \ Cov(V_i(t_k), V_i(t_{k'})) = e^{-2|t_k - t_{k'}|} \ .$$

La fonction moyenne de $Z(.)$ *est nulle* H_0 *et vérifie sous* $H_{a,t^\star}$*:*

$$m_{Z,t^\star}(t) = \frac{\sqrt{\mathcal{A}_1} \ \xi_1(t) m_{V_1,t^\star}(t) \ + \ \sqrt{\mathcal{A}_2} \ \xi_2(t) m_{V_2,t^\star}(t)}{\sqrt{\mathcal{A}_1 \ \xi_1^2(t) + \mathcal{A}_2 \ \xi_2^2(t)}}.$$

Les $\alpha_i(t)$*,* $\beta_i(t)$ *et* $\xi_i(t)$ *des fonctions connues et*

$$m_{V_i,t^\star}(t) = \left\{ \alpha_i(t) \ m_{V_i,t^\star}(t^{\ell,i}) + \beta_i(t) \ m_{V_i,t^\star}(t^{r,i}) \right\} / \xi_i(t)$$
$$\forall t_k \in \mathbb{T}_K^i \ \ m_{V_i,t^\star}(t_k) = \frac{\sqrt{\mathcal{A}_i}}{\sigma^2} \sum_{s=1}^{m} a_s \ e^{-2|t_s^\star - t_k|} \ .$$

D'après le théorème précédent, en discrétisant le processus de score sur la position des marqueurs, nous avons quand n est grand:

$$\vec{S}_n = \vec{m}_{Z,t^\star} \ + \ \vec{\varepsilon} \ + \ o_P(1)$$

où $\vec{S}_n = (S_n(t_1) \ , \ S_n(t_2) \ , \ ... \ , \ S_n(t_K))' \ , \ \vec{m}_{Z,t^\star} = (m_{Z,t^\star}(t_1) \ , \ m_{Z,t^\star}(t_2) \ , \ ..., m_{Z,t^\star}(t_K))'$ et $\vec{\varepsilon} \sim N(0, \Sigma)$ avec $\Sigma_{kk'} = \text{Cov}\left(Z(t_k), Z(t_{k'})\right)$.

Dans ce qui suit, on se place en déséquilibre de liaison complet, i.e. les m QTLs sont localisés sur certains marqueurs. Ainsi, on recherchera des QTLs seulement aux

emplacements des marqueurs. Grace aux spécificités de la fonction moyenne du processus et après avoir décorrélé les composantes de $\vec{S}_n$ en considérant la décomposition de Cholesky $\Sigma = AA'$, on obtient la relation suivante :

$$A^{-1}\vec{S}_n = A'\left(\Delta_1\ ,\ \ldots\ ,\ \Delta_K\right)' \ +\ A^{-1}\vec{\varepsilon}\ +\ o_P(1) \tag{8}$$

$$\text{où}\ \ \Delta_k = \begin{cases} 0 & \text{si} \quad t_k \notin \{t_1^\star, \ldots, t_m^\star\} \\ \frac{a_s\sqrt{\mathcal{B}}}{\sigma^2} & \text{si } t_k \in \{t_1^\star, \ldots, t_m^\star\} \cap \mathbb{T}_K^2 \text{ avec } s \text{ l'indice tel que } t_s^\star = t_k. \\ \frac{a_s\sqrt{\mathcal{A}_1+\mathcal{A}_2\xi_2^2(t_k)}}{\sigma^2} & \text{si } t_k \in \{t_1^\star, \ldots, t_m^\star\} \cap \mathbb{T}_K^1\backslash\mathbb{T}_K^2 \text{ avec } s \text{ l'indice tel que } t_s^\star = t_k. \end{cases}$$

Les QTLs placés sur les marqueurs de la carte discrète se voient amplifiés d'un facteur $\sqrt{\mathcal{B}}/\sigma$ et d'un facteur $\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)}/\sigma$ sur la carte dense. Ainsi, contrairement au génotypage sélectif classique où tous les loci disposent du même coefficient multiplicatif (Rabier et Delmas, 2021), on voit que désormais les facteurs multiplicatifs diffèrent en fonction de l'appartenance des loci aux 2 cartes.

Enfin, afin de trouver les Δ_k non nuls, une méthode naturelle est d'utiliser la régression pénalisée L1, appelée Lasso. Ainsi, en notant $\Delta := (\Delta_1, ..., \Delta_K)'$, l'estimateur AdaptSgeno-Lasso s'écrit

$$\hat{\Delta}_{\text{AdaptSgenoLasso}}(\lambda, \alpha) = \arg\min_{\Delta}\left(\left\|A^{-1}\vec{S}_n - A'\Delta\right\|_2^2 + \lambda\left\|\Delta\right\|_1\right)\ . \tag{9}$$

Notons que l'on pourrait également accorder plus d'importance aux loci gènes majeurs en couplant notre proche avec l'Adaptative Lasso. Cela correspondrait à imposer une pénalité dans la formule (9) de la forme $\|W'\Delta\|_1$ avec des poids W_k égaux à $1/\sqrt{\mathcal{B}}$ sur la carte $\mathbb{T}_K^2$ (i.e. gènes majeurs) et $1/\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)}$ sur la carte $\mathbb{T}_K^1\backslash\mathbb{T}_K^2$.

Bibliographie

[1] Cierco C. (1998), Asymptotic distribution of the maximum likelihood ratio test for gene detection, *Statistics*, 31 261-285.

[2] Chang, M.N., Wu, R., Wu, S.S., and Casella, G. (2009), Score statistics for mapping quantitative trait loci, *Stat. Appl. Genet. Mol. Biol.*, 8(1) 16.

[3] Fan, J., Li, Q., Wang, Y. (2017), Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1) (2017), pp. 247-265.

[4] Lebowitz RJ, Soller M, Beckmann, J.S. (1987), Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines, *Theor. Appl. Genet.*, 73 556-562.

[5] Rabier C-E, Delmas C (2021): The SgenoLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection, *Statistics*, DOI: 10.1080/02331888.2021.1881785.

NOUVELLE SUBVENTION POUR LA FORMATION PROFESSIONNELLE SUPÉRIEURE : PROFIL DES PREMIERS BÉNÉFICIAIRES

Anne Renaud[1] & Réjane Deppierraz[2] & Nicole Schöbi[3] & Melanie Stutz[4]

[1234] *Office fédéral de la statistique, Espace de l'Europe 10, CH-2010 Neuchâtel, Suisse*
[1]*anne.renaud@bfs.admin.ch;* [2]*rejane.deppierraz@bfs.admin.ch;*
[3]*nicole.schoebi@bfs.admin.ch;*[4]*melanie.stutz@bfs.admin.ch*

Résumé. Un nouveau système de subventions a été mis en place en Suisse en 2018 afin de soutenir financièrement les candidat/es aux examens fédéraux (formation professionnelle supérieure). Cette même année, 23'000 personnes se sont présentées aux examens fédéraux et un quart d'entre elles ont demandé la subvention. Cette étude a pour objectif d'examiner les caractéristiques de ses premiers bénéficiaires. L'impact sur la demande de subvention de caractères démographiques (sexe, âge, nationalité), de facteurs contextuels (région linguistique, type de commune) et de spécificités liées à la formation (par ex. domaine de formation et réussite à l'examen) a été évalué à l'aide de modèles de régression logistique. Les résultats ont montré que les variables liées à la formation étaient les plus significatives. La relativement faible performance des modèles estimés suggère cependant que l'intégration d'indicateurs sur la structure de la formation dans les différentes associations professionnelles devrait être évaluée pour mieux comprendre le profil des bénéficiaires.

Mots-clés. Formation professionnelle supérieure, nouvelle subvention, régression logistique

Abstract. A new subsidy system was introduced in Switzerland in 2018 providing financial support to candidates for federal (advanced) professional examinations (professional education at tertiary level). That same year, 23 000 people took the federal exams, and a quarter of them applied for the subsidy. The aim of this study is to examine the characteristics of the first beneficiaries. Logistic regression modelling was used to assess the impact of demographic characteristics (gender, age, nationality), contextual factors (language region, type of municipality) and training features (e.g. training fields and examination success) on the subsidy demand. The results showed that variables related to training were the most significant. The relatively poor performance of the estimated models suggests, however, that indicators on the structure of training in the different trade associations should be integrated in order to better understand the profile of beneficiaries.

Keywords. Professional education, new subsidy, logistic regression

1. Introduction

Dans le système éducatif suisse, la formation professionnelle supérieure (FPS) est l'un des piliers du degré tertiaire au même titre que la formation délivrée dans les hautes écoles (voir annexe). Elle permet d'acquérir des qualifications en vue d'exercer des activités professionnelles complexes avec des responsabilités élevées. Dans le cadre du projet stratégique destiné à renforcer la FPS, la Confédération a mis en place un nouveau régime de financement axé sur la personne. Depuis 2018, les candidat/es aux examens fédéraux domiciliés en Suisse peuvent demander le remboursement de la moitié des frais déboursés pour les cours préparatoires commencés après le 1er janvier 2017, quel que soit le résultat obtenu à l'examen. En 2018, 23'000 personnes se sont présentées aux examens

fédéraux et près de 5'300 d'entre elles se sont vu rembourser la moitié des frais de cours préparatoire (23% des candidat/es; OFS, 2020a ; OFS, 2020b). Il s'agit des premiers bénéficiaires du nouveau système de financement axé sur la personne.

De nombreuses études en éducation analysent les différences susceptibles de refléter des déséquilibres ou des inégalités mais peu sont disponibles pour les examens fédéraux ou la FPS. Néanmoins, on peut citer ici quelques éléments. Tout d'abord, plus d'hommes que de femmes suivent une FPS après une formation professionnelle initiale (32% et 23% respectivement; OFS, 2020e). Cette différence se reflète aussi dans la part de population résidante de plus de 25 ans ayant achevé une FPS (18% des hommes et 11% des femmes ; OFS, 2020d). Ensuite, des différences ont été également observées entre régions linguistiques. La part de la population résidante de plus de 25 ans ayant achevé une FPS était plus élevée dans la région de langue allemande (16%) comparée aux régions francophone et italophone (11-12%) (OFS, 2020d).

La présente étude s'intéresse aux caractéristiques des premiers bénéficiaires de la subvention. Elle vise à répondre à la question d'étude suivante : Quels sont les facteurs qui caractérisent les candidat/es 2018 qui ont déposé une demande (et obtenu une subvention) par rapport aux autres? Sachant que toute personne satisfaisant les critères administratifs d'octroi reçoit la subvention, on tente de déterminer par une approche multivariée si la disposition d'une personne à faire une demande de subvention est influencée par des facteurs liés à la formation (p. ex. domaine de formation), par ses caractéristiques démographiques (p. ex. sexe et âge) ou encore par des facteurs contextuels (p. ex. région linguistique). Le nouveau système de financement se trouve dans sa période d'introduction. L'analyse du profil des premiers bénéficiaires pourrait identifier des déséquilibres dans cette période particulière entre les domaines de formation ou entre des groupes de personnes selon des caractéristiques contextuelles ou démographiques.

2. Méthode

Deux jeux de données principaux ont formé la base de l'analyse: l'ensemble des candidat/es aux examens fédéraux en 2018 (examens professionnels et professionnels supérieurs, domicile en Suisse) et l'ensemble des bénéficiaires de subventions octroyées jusqu'à fin 2019 pour ces mêmes examens. Les données sur les candidat/es étaient relevées par la statistique des diplômes 2018 (SBA). Les données sur les bénéficiaires provenaient de la statistique du financement axé sur la personne en formation professionnelle supérieure (aHBB, demandes de subventions acceptées, après l'examen). Un appariement a été effectué entre la liste des candidat/es 2018 et celle des bénéficiaires afin de déterminer, dans chaque cas, si une subvention a été perçue. Une correspondance a été trouvée pour 99% des subventions reçues en 2018 et 96% de celles reçues en 2019 pour des examens effectués en 2018.

Les candidat/es aux examens fédéraux de métiers tels que policier/ère, agent/e de détention ou garde-frontière n'ont pas droit à la subvention fédérale car leur formation est subventionnée par les cantons. Ces environ 1'100 candidat/es ont donc été exclus de la population d'analyse. Par contre, une personne était présente plusieurs fois si elle pouvait demander plusieurs subventions (p. ex. candidat à deux brevets différents). Le jeu de données pour l'analyse comportait ainsi 21'839 candidat/es aux examens fédéraux en 2018.

Pour répondre à la question d'étude, l'analyse a été organisée en deux parties. Premièrement, une analyse descriptive des données a permis d'observer les caractéristiques des candidat/es et la proportion d'entre elles/eux ayant obtenu une subvention dans différents sous-groupes (taux de subvention). Les variables sélectionnées pour l'analyse ont été rassemblées en trois types. Le premier correspondait aux caractéristiques démographiques (sexe, âge et nationalité). Le second englobait les variables contextuelles (région linguistique et type de commune de domicile). Le dernier reflétait le cadre de la formation (domaine de formation selon la classification internationale ISCED-F 2013, examen professionnel ou examen professionnel supérieur ainsi que résultat à l'examen). Ce choix s'est porté sur les variables disponibles dans la statistique des diplômes et

usuelles dans la recherche sur l'éducation ; voir p. ex. CSRE (2018) et Backes-Gellner *et al.* (2020). Deuxièmement, des modèles de régression logistique ont permis d'évaluer le lien entre la probabilité d'avoir fait une demande (variable dichotomique, 1 pour demande de subvention) et les différentes variables explicatives. La relation entre les variables explicatives et la variable à expliquer a été testée par le chi-carré (test d'indépendance). La multicolinéarité entre les variables explicatives a été également évaluée avant de sélectionner les variables intégrées dans le modèle (*variance inflation factor VIF* < 10 ; Tuffery, 2010). Différents modèles ont été définis. Leur qualité globale a été mesurée par les critères d'information *AIC* et *BIC*, le *likelihood ratio* test et le pseudo-R^2 de *McFadden ; c*elle des coefficients par le test de *Wald* (Peng *et al*, 2002 et Tuffery, 2010). Dans un objectif de parcimonie, les variables non significatives ont été exclues des modèles (*stepwise* avec l'*AIC*). L'objectif de l'étude était exploratoire. La précision (taux de vrais positifs et vrais négatifs), la spécificité (1 - taux de faux positifs) et la sensibilité (taux de vrais positifs) d'une prévision basée sur les modèles ont été cependant évaluées afin de mieux comprendre les caractéristiques des modèles estimés (table de classification sur la base des probabilités prédites). L'analyse a été effectuée dans R. Les valeurs *p* inférieures à 5% étaient considérées comme significatives.

3. Résultats

Les 21'839 candidat/es aux examens fédéraux étaient principalement des hommes (63%) et 44% avaient moins de 30 ans (cf. Tableau 1). La grande majorité des examens étaient des examens professionnels (83%). Le taux de subvention global était de 24% et variait entre les sous-groupes. On note par exemple que les femmes présentaient un taux plus élevé que les hommes (27% et 22% respectivement). Il en allait de même pour les candidat/es aux examen professionnels supérieurs comparés aux candidat/es aux examens professionnels (29% et 23%). Des différences étaient également observées entre les domaines de formation, avec notamment un taux de 9% pour Santé et protection sociale et 41% pour Marketing et publicité. Toutes les variables explicatives étaient significativement dépendantes de la variable à expliquer (« demande de subvention », test du chi-carré, $p < 0.01$). Aucun problème de multicolinéarité n'a été détecté parmi les variables explicatives sélectionnées ($VIF \leq 1.6$).

Un modèle logistique de base a permis de faire le lien entre la demande de subvention (variable dichotomique) et les variables explicatives dans leur ensemble (cf. Tableau 2). Les trois facteurs liés à la formation étaient les plus significatifs. Venaient ensuite la région linguistique, le sexe, le type de commune, la classe d'âge et enfin la nationalité. Le Tableau 2 indique les *odds ratio (OR)* estimés. Parmi les domaines de formation, Santé et protection sociale ou Technologie et ingénierie montraient par exemple une propension aux demandes inférieure au groupe de référence (Gestion et administration) ($OR<1$) alors que l'inverse était observé pour Marketing et publicité ainsi que pour Finance, banque et assurances ($OR>1$). Les personnes se présentant aux examens professionnels supérieurs, celles domiciliées dans des communes intermédiaires ainsi que les femmes étaient également plus susceptibles de demander une subvention que les catégories de référence ($OR>1$). A l'inverse, les personnes ayant échoué à l'examen, celles de la région italophone ou encore celles de plus de 50 ans en avaient moins déposé ($OR<1$).

Le modèle de base était significatif pour expliquer la variable « demande de subvention=1 » (*likelihood ratio* test, $p < 0.001$). Par contre sa capacité d'explication était plutôt faible. Le pseudo R^2 ne valait en effet que 0.06. Un nouveau modèle a donc été estimé en ajoutant des interactions entre paires de variables explicatives. Ces interactions ont été testées puis un large choix ajouté au modèle. Après une sélection du type *stepwise* avec l'*AIC*, dix interactions ont été finalement conservées. Il s'agissait notamment du domaine de formation croisé avec le type d'examen, le sexe, la région linguistique et le résultat à l'examen. Le Tableau 3 permet de comparer le modèle de base et le modèle « étendu ». Ce dernier était significativement meilleur que le modèle de base (*likelihood ratio* test, $p < 0.001$). Par contre, malgré l'ajout de nombreuses interactions, le pseudo R^2 restait faible (0.09). La prédiction basée sur le modèle correspondait à la vraie valeur dans plus

de trois-quarts des cas (précision = 77%). La part de prédiction correcte était élevée parmi les candidat/es n'ayant pas fait de demande (spécificité = 97%) et plus basse pour les candidat/es avec demande (sensibilité = 12%).

Tableau 1 : Candidat/es et subventions pour les examens 2018

Facteur	Catégorie	Nombre de candidat/es	Part de candidat/es [%]	Nombre de subventions	Taux de subventions [%]
Total		21839	100%	5206	24%
Facteurs démographiques					
Sexe	homme	13767	63%	3034	22%
	femme	8072	37%	2172	27%
Classe d'âge	-29	9640	44%	2317	24%
	30-39	7635	35%	1906	25%
	40-49	3178	15%	733	23%
	50+	1385	6%	250	18%
	missing	1	0%	0	0%
Nationalité	suisse	19316	88%	4664	24%
	étrangère	2421	11%	524	22%
	missing	102	0%	18	18%
Facteurs contextuels					
Région linguistique	Allemand	18171	83%	4363	24%
	Français	3271	15%	793	24%
	Italien	397	2%	50	13%
Type de commune	Urbain	12944	59%	3175	25%
	Intermédiaire	4925	23%	1200	24%
	Rural	3970	18%	831	21%
Facteurs liés à la formation					
Type d'examen	Examen prof.	18029	83%	4108	23%
	Examen prof. supérieur	3810	17%	1098	29%
Domaine de formation	Arts, lettres, langues et sci. sociales	880	4%	239	27%
	Comptabilité et fiscalité	1959	9%	642	33%
	Finance, banque et assurances	1785	8%	666	37%
	Gestion et administration	3237	15%	980	30%
	Marketing et publicité	1007	5%	411	41%
	Vente en gros et au détail	3078	14%	745	24%
	Sciences nat., agri. et sc. vété	960	4%	97	10%
	Technologie et ingénierie	4991	23%	872	17%
	Santé et protection sociale	1615	7%	139	9%
	Services	2327	11%	415	18%
Résultat à l'examen	réussi	15916	73%	4075	26%
	pas réussi	5923	27%	1131	19%

Note : subventions jusqu'à fin 2019. La région de langue romanche est incluse dans la région de langue allemande.

Tableau 2 : Modèle de base pour la demande de subvention pour les examens 2018.

Facteur	Catégorie	Coefficient	Ecart-type	*Wald's* X^2	*Odds ratio (OR)*	Intervalle 95% *OR*
Constante		-0.7814	0.0503	-15.533***	0.46	(0.41-0.51)
Sexe (réf.: homme)	femme	0.1233	0.0372	3.318***	1.13	(1.05-1.22)
Classe d'âge	30-39	-0.0106	0.0378	-0.28	0.99	(0.92-1.07)
(réf.: -29)	40-49	-0.0561	0.0515	-1.09	0.95	(0.85-1.05)
	50+	-0.1958	0.0789	-2.481*	0.82	(0.70-0.96)
Nationalité (réf.: suisse)	étrangère	-0.0827	0.0543	-1.522	0.92	(0.83-1.02)
Région linguistique	Français	0.0161	0.0467	0.345	1.02	(0.93-1.11)
(réf.: Allemand)	Italien	-0.7614	0.1556	-4.893***	0.47	(0.34-0.63)
Type de commune	Intermédiaire	0.0842	0.0406	2.073*	1.09	(1.00-1.18)
(réf. Urbain)	Rural	-0.0385	0.0465	-0.828	0.96	(0.88-1.05)
Type d'examen (réf.: prof)	Examen prof. supérieur	0.5204	0.0445	11.69***	1.68	(1.54-1.84)
Domaine de formation	Arts, lettres, langues, sc.soc.	-0.1835	0.0866	-2.12*	0.83	(0.70-0.99)
(réf.: Gestion et administration)	Comptabilité et fiscalité	-0.0133	0.0651	-0.204	0.99	(0.87-1.12)
	Finance, banque et assurance	0.2850	0.0631	4.519***	1.33	(1.17-1.50)
	Marketing et publicité	0.3783	0.0760	4.977***	1.46	(1.26-1.69)
	Vente en gros et au détail	-0.3759	0.0580	-6.486***	0.69	(0.61-0.77)
	Sciences nat., agri., sc. vét.	-1.4436	0.1170	-12.334***	0.24	(0.19-0.30)
	Technologie et ingénierie	-0.7810	0.0576	-13.564***	0.46	(0.41-0.51)
	Santé et protection sociale	-1.6919	0.1008	-16.794***	0.18	(0.15-0.22)
	Services	-0.7461	0.0676	-11.035***	0.47	(0.42-0.54)
Résultat à l'examen (réf.: réussi)	pas réussi	-0.4197	0.0394	-10.653***	0.66	(0.61-0.71)

Note : ***p < 0.001; ** p < 0.01; * p < 0.05

Tableau 3 : Comparaison des modèles sans et avec interactions pour la demande de subvention.

	Modèle sans interaction (*base*)	Modèle avec interactions (*étendu*)
Facteurs de base	Sexe, classe d'âge, nationalité, région linguistique, type de commune, type d'examen, domaine de formation et résultats à l'examen	Sexe, classe d'âge, nationalité, région linguistique, type de commune, type d'examen, domaine de formation et résultats à l'examen
Interactions	-	Domaine:type d'examen, Domaine:sexe, Domaine:région ling., Classe d'âge:type examen, Domaine:rés. examen, Sexe:rés. examen, Type commune:rés. examen, Région ling. :type commune, Sexe:type examen, Sexe:région ling.
AIC/BIC	22616/22784	21941/22573
Log Likelihood	-11286	-10891
Likelihood ratio test (df)	1316.5 (20)***	2106.3 (78)***
Pseudo R^2 (*McFadden*)	0.06	0.09
Différence entre les deux modèles *likelihood ratio* test (df)		789.89 (58)***

Note : ***p < 0.001; ** p < 0.01; * p < 0.05

4. Discussion

Cette étude avait pour objectif de caractériser les premiers bénéficiaires de la nouvelle subvention axée sur la personne en formation professionnelle supérieure. Les résultats complètent ceux obtenus par OFS (2020a, 2020b).

Les facteurs liés à la formation (domaine de formation, type d'examen et résultats à l'examen) ont été identifiés comme ayant le plus d'impact sur la demande de subvention. Les différences entre les domaines de formation pourraient être liées à la structure de la formation et au degré d'adaptation au nouveau système. En effet, une partie des candidat/es n'ont pas pu faire de demande car ils n'avaient fréquenté que des cours commencés avant 2017. On pense notamment au domaine Santé et protection sociale pour lequel les formations sont plutôt longues par rapport aux autres domaines (OFS, 2020c). Les différences entre domaines pourraient se réduire à l'avenir. Le peu de demandes de celles/ceux ayant échoué à l'examen pourrait provenir d'un problème dans la communication (la subvention peut aussi être demandée si l'examen n'est pas réussi) ou d'une tendance à demander la subvention après avoir répété l'examen.

Les facteurs contextuels et démographiques étaient également significatifs mais de moindre importance. Des explications sur le peu de demandes parmi les candidat/es de la région italophone pourraient être explorées du côté de spécificités régionales (communication incluse). La dynamique

dans le choix de la FPS et celle dans la demande de subvention semblent différentes. En effet, la population de la région italophone opte moins pour la FPS que la région germanophone (OFS, 2020d) et ses candidat/es ont aussi moins profité de la subvention. Quant aux femmes, elles choisissent moins la FPS que les hommes (OFS, 2020d ; OFS, 2020e) mais ont plus profité de la subvention.

La qualité des modèles laisse supposer que des variables manquent pour expliquer la demande de subvention. La grande majorité des candidat/es étant employé/es (env. 96% ; OFS, 2020c), des facteurs explicatifs pourraient être considérés au niveau des associations professionnelles (ex. durée et structure des cours, communication sur les subventions) et des entreprises (ex. types de convention de formation entre employé/es et employeurs).

Certaines limites de l'étude méritent l'attention. Premièrement, l'ensemble des candidat/es aux examens 2018 ayant fait une demande de subvention ne sera connu que fin 2020 (délai de 2 ans). Le nombre de demandes tardives devraient cependant être petit. Deuxièmement, les candidat/es ne satisfaisant pas aux critères pour faire une demande n'ont pas pu tous être identifiés a priori. Parmi eux se trouvaient ceux qui n'ont pas suivi de cours préparatoires (env. 5% ; OFS, 2020c) et ceux n'ayant fréquenté que des cours hors de la liste (p. ex. cours commencés avant 2017). Ces éléments sont liés à la structure de la formation.

L'analyse pourrait être réitérée à l'avenir afin de voir si les déséquilibres perdurent. De plus, une analyse qualitative avec des expert/es ou des informations sur la structure de la formation permettraient d'affiner les connaissances sur le profil des bénéficiaires de la nouvelle subvention.

Bibliographie

[1] Backes-Gellner, U., Renold, U. & Wolter, S. C (2020). *Economics and governance of vocational and professional education and training (including apprenticeship): theoretical and empirical results for researchers and educational policy leaders*. Bern: hep Verlag.

[2] Centre suisse de coordination pour la recherche en éducation [CSRE] (2018). *L'éducation en Suisse – rapport 2018*, Aarau

[3] Office fédéral de la statistique [OFS] (2020a). *Subventions fédérales pour cours préparatoires aux examens fédéraux. Formation professionnelle supérieure – financement axé sur la personne 2018*, Neuchâtel

[4] Office fédéral de la statistique [OFS] (2020b). *Subventions fédérales pour cours préparatoires aux examens fédéraux. Evolution des subventions entre 2018 et 2019*, Neuchâtel

[5] Office fédéral de la statistique [OFS] (2020c). *Conditions de formation des candidats aux examens de la formation professionnelle supérieure. Résultats de l'enquête sur la formation professionnelle supérieure 2019*, Neuchâtel

[6] Office fédéral de la statistique [OFS] (2020d). Formation achevée la plus élevée en 2018 https://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/niveau-formation/niveau-formation-regional.html, consulté le 15.2.2021, Neuchâtel

[7] Office fédéral de la statistique [OFS] (2020e). Indicateurs de la formation. https://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/indicateurs-formation.html, consulté le 23.2.2021, Neuchâtel

[8] Office fédéral de la statistique [OFS] (2020f). Diplômes. https://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/diplomes.html, consulté le 23.2.2021, Neuchâtel

[8] Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research,* 96:1, 3-14

[9] Tuffery, S. (2010). *Data mining et statistique décisionnelle*. Editions Technip, Paris

Annexe : la formation professionnelle supérieure en Suisse

La *formation professionnelle supérieure* (FPS) fait partie du degré tertiaire du système éducatif suisse au même titre que les hautes écoles. Elle combine l'enseignement et la pratique professionnelle selon un système dual - formation en entreprise et en école - pour plus de 600 professions.

Un titre de la FPS s'acquiert soit par un *examen fédéral* (examen professionnel ou examen professionnel supérieur) soit par une formation en école supérieure. L'examen professionnel est sanctionné par un *brevet fédéral* (p. ex. Spécialiste en finance et comptabilité avec brevet fédéral) et l'examen professionnel supérieur par un *diplôme fédéral* (p. ex. Installateur/trice –électricien/ne diplômé/e). Une formation en école supérieure débouche sur un diplôme d'une école supérieure (p. ex. Infirmier/ère diplômé/e ES). Comme illustré sur la figure 1, la FPS a délivré un total d'environ 26'500 diplômes en 2018. Ce nombre correspond à un tiers des 78'100 diplômes du degré tertiaire. Les brevets et diplômes fédéraux représentent env. deux tiers des diplômes de la FPS, et les diplômes des écoles supérieures un tiers. Le nombre total de candidat/es aux examens de la FPS était de l'ordre de 33'200, dont 23'500 pour les examens fédéraux (OFS, 2020f).

La FPS relève de la compétence de la Confédération, en partenariat avec les cantons et les organisations du monde du travail. Les cours préparatoires aux examens fédéraux sont pour l'essentiel financés par le secteur privé (entreprises et personnes en formation) alors que les écoles supérieures sont principalement financées par les pouvoirs publics (cantons et Confédération). La préparation aux examens fédéraux dure en moyenne environ une année et demie à deux ans avec de grandes différences entre les domaines de formation. De même, les coûts de formation et le soutien des employeurs varie sensiblement entre les domaines de formation et au sein de ceux-ci (OFS, 2020c).

En 2018, la Confédération a renforcé son soutien à la formation professionnelle supérieure. Depuis cette année, les candidat/es aux examens fédéraux se voient rembourser la moitié des frais déboursés pour les cours préparatoires s'ils se sont présentés à un examen fédéral et quel que soit le résultat à l'examen (*financement axé sur la personne*). La demande peut être déposée jusqu'à deux ans après la notification du résultat de l'examen. Les cours peuvent être cumulés mais doivent tous avoir commencé après le 1er janvier 2017, avoir été payés par la personne et être présent sur la « liste des cours préparatoires » de l'année concernée (env. 5'000 cours). On notera que des cours subventionnés par les cantons sont exclus de la liste des cours (p. ex. formation pour la police).

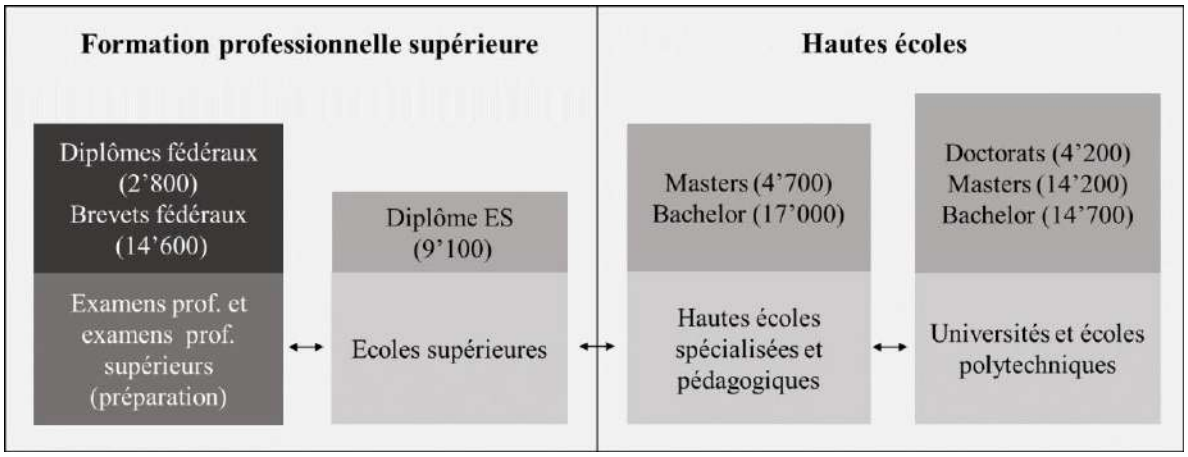


Figure 1 : Le degré tertiaire du système de la formation en Suisse et nombre de diplômes en 2018.

Non-asymptotic statistical test of the covariance matrix rank of a 2-dimensional SDE

Anna Melnykova[1], Patricia Reynaud-Bouret[2], Adeline Samson[3]

[1] *Grenoble INP, LJK UMR-CNRS 5224,*
E-mail: anna.melnykova@univ-grenoble-alpes.fr
[2] *Université de Nice Sophia-Antipolis, UMR-CNRS 7351*
E-mail: reynaudb@unice.fr
[3] *Université Grenoble Alpes, LJK UMR-CNRS 5224,*
E-mail: adeline.leclercq-samson@univ-grenoble-alpes.fr

Résumé. Le but de ce travail est de développer une procédure de test non-asymptotique qui détermine le rang de la matrice de diffusion dans un processus stochastique bi-dimensionnel à partir d'observations discrètes de ce processus sur un intervalle de temps fixe $[0, T]$ échantillonné avec un pas de temps Δ. Tout d'abord, nous construisons les statistiques principales du test. Ces statistiques sont définies à l'aide de déterminants de matrices aléatoires, comme proposé dans Jacod & Podolskij (2013). Nous montrons que les performances du test basé sur ces statistiques sont limitées dans un cadre non asymptotique, lorsque Δ est fixe. Ensuite, nous montrons comment les performances du test peuvent être améliorées en centrant les incréments du processus autour de leur valeur attendue (terme de dérive). Enfin, nous déduisons la distribution des statistiques centrées et montrons sous quelles conditions les erreurs de type I et de type II du test peuvent être contrôlées.

Mots-clés. Tests statistiques, statistique de processus, modèle neuronal FitzHugh-Nagumo, statistique non-asymptotique, matrices aléatoires

Abstract. The aim of this work is to develop a testing procedure which determines the rank of the diffusion matrix in a two-dimensional stochastic process from discrete observations of this process on a fixed time interval $[0, T]$ sampled with a fixed time step Δ. First, we construct the main statistics of the test, given by a random matrix determinant, as proposed in Jacod & Podolskij (2013). We show that the performance of the test based on this statistics is limited in a non-asymptotic setting, when Δ is fixed. Then, we show how the performance of the test can be improved by centering the increments of the process around their expect value (drift term). Finally, we derive the distribution of the centered statistics and show under which conditions the Type I and Type II errors of the test can be controlled.

Keywords. Statistical tests, statistics of the processes, neuronal FitzHugh-Nagumo model, non-asymptotic statistics, random matrices

1 Introduction

Consider a 2-dimensional process $X = (X^1, X^2)^T$, defined by the solution of:

$$dX_t = b_t dt + \sigma dW_t, \tag{1}$$

where $b_t = (b_t^1, b_t^2)^T$ is a non-constant drift vector and σ is a diagonal diffusion matrix with constant coefficients σ_1 and σ_2 on the main diagonal, W is a 2-dimensional Brownian motion. The goal is to construct the test of the following hypothesis:

$$H_0 : \sigma_1^2 \sigma_2^2 = \delta$$
$$H_1 : \sigma_1^2 \sigma_2^2 \geq \delta,$$

where δ is some chosen "sensitivity" threshold. H_0 and H_1 correspond roughly to the case of the covariance matrix being of a full rank (i.e., (1) is *elliptic*) or not (*hypoelliptic* or deterministic), as δ can be arbitrarily close to 0.

2 Existing works

In Jacod et al. (2008) and Jacod & Podolskij (2013) the principal statistics, determining the "noisiness" in the system, are defined as a sequence of the random matrix determinants

$$S = \frac{1}{n} \sum_{i=1}^{n} \det \Phi_i^2, \tag{2}$$

where the matrices are constructed on the non-overlapping increments of the process X as follows:

$$\Phi_i = \frac{1}{\Delta} \begin{pmatrix} X^k_{(2i+1)\Delta} - X^k_{2i\Delta} & X^k_{(2i+2)\Delta} - X^k_{(2i+1)\Delta} \\ X^k_{(2i+1)\Delta} - X^k_{2i\Delta} & X^k_{(2i+2)\Delta} - X^k_{(2i+1)\Delta} \end{pmatrix}. \tag{3}$$

In Jacod & Podolskij (2013) it is proven that asymptotically, as $\Delta \to 0$, the statistics S allow for identifying the rank of the covariance matrix of the process X. In practice, however, the asymptotic setting is unreachable, since the time step Δ is usually fixed to a strictly positive value. In addition, if the variance coefficients are not zero, but very different (for example, $\sigma_2 = 10\sigma_1$), the asymptotic test will fail to identify the difference between the elliptic and hypoelliptic case.

3 Our approach and principal results

To overcome the difficulties described above, we propose to consider the adapted statistics, defined as in (2), but with the matrices constructed as follows:

$$\dot{\Phi}_i = \frac{1}{\Delta} \begin{pmatrix} X^k_{(2i+1)\Delta} - X^k_{2i\Delta} - \int_{2i\Delta}^{(2i+1)\Delta} b_s ds & X^k_{(2i+2)\Delta} - X^k_{(2i+1)\Delta} - \int_{(2i+1)\Delta}^{(2i+2)\Delta} b_s ds \\ X^k_{(2i+1)\Delta} - X^k_{2i\Delta} - \int_{2i\Delta}^{(2i+1)\Delta} b_s ds & X^k_{(2i+2)\Delta} - X^k_{(2i+1)\Delta} - \int_{(2i+1)\Delta}^{(2i+2)\Delta} b_s ds \end{pmatrix}. \tag{4}$$

In other words, we center each entry of the matrix Φ_i around its expected value, so that we obtain a new matrix with centered Gaussian entries. In practice, the drift is often unknown. However, it can be estimated either parametrically or non-parametrically. In our experiments, conducted on FitzHugh-Nagumo neuronal model, we estimate it with a least squares parameter estimation. In theoretical results, we assume the drift to be known.

Our first result gives the cumulative distribution function of each summand of the statistics S.

Proposition 1. *Denote $\dot{s}_i := \det \dot{\Phi}_i^2$, where $\dot{\Phi}_i^2$ is defined in (4). Then, the following holds for all i:*

$$\mathbf{P}\left(\dot{s}_i \leq x\right) = 1 - \left(\sqrt{\frac{x}{\sigma_1^2\sigma_2^2}} + 1\right) e^{-\sqrt{\frac{x}{\sigma_1^2\sigma_2^2}}}$$

Then, basing on this result we obtain the quantile for the test, defined in Section 1:

Theorem 1. *Under H_0 the following bound holds:*

$$\mathbf{P}_0\left(\dot{S} \geq \delta\left(1 + W\left(-\frac{\alpha^{1/n}}{e}\right)\right)^2\right) \leq \alpha,$$

where W denotes Lambert W function.

Theorem 1 gives us the following rejection rule of the test:

$$H_0 \text{ is rejected if } \dot{S} \geq z_\alpha,$$

where z_α is defined as

$$z_\alpha := \delta\left(1 + W\left(-\frac{\alpha^{1/n}}{e}\right)\right)^2. \tag{5}$$

Finally, we obtain the control on the power of the test with the following result:

Theorem 2. *For fixed levels of Type I and Type II risks α and β respectively and if*

$$\sigma_1^2\sigma_2^2 \geq \delta\left(\frac{1 + W\left(-\frac{\alpha^{1/n}}{e}\right)}{1 + W\left(-\frac{(1-\beta)^{1/n}}{e}\right)}\right)^2,$$

the following inequality holds under H_1:

$$\mathbf{P}_1\left(\dot{S} \leq z_\alpha\right) \leq \beta,$$

We conclude with a numerical study, conducted on neuronal FitzHugh-Nagumo model (FitzHugh, 1961). We couple the test with a parametric estimator, which gives as an accurate approximation of the drift term.

Bibliographie

Fitzhugh, R. (1961). *Impulses and physiological states in theoretical models of nerve membrane.* Biophysical Journal, 1(6):445–466.
Jacod, J., Lejay, A., Talay, D., et al. (2008). *Estimation of the Brownian dimension of a continuous Itô process.* Bernoulli, 14(2):469–498.
Jacod, J. and Podolskij, M. (2013). *A test for the rank of the volatility process: the random perturbation approach.* The Annals of Statistics, 41(5):2391–2427.

A Bound on the expected runtime of the pDPA algorithm for multiple changepoint detection

Guillem Rigaill [1,2,3]

[1] *Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRAE, Université Paris-Sud, Université Evry, Université Paris Saclay, Batiment 630, 91405 Orsay, France*
[2] *Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405, Orsay, France*
[3] *Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Evry Val d'Essonne, 23 boulevard de France, 91037 Evry, France*

Résumé. L'algorithme pDPA estime simultanément la position de K ruptures dans un signal univarié en maximisant une vraisemblance. Le pDPA a une complexité au pire quadratique en la taille des données mais en pratique elle est souvent log-linéaire. Dans ce travail nous étudions l'espérance du temps de calcul du pDPA pour $K = 1$ rupture : pDPA(1). Nous démontrons que pour des données $(Y_1, \cdots, Y_n)$ obtenues comme la somme d'un signal constant par morceaux $(\mu_1, \cdots, \mu_n)$ en $D + 1$ morceaux et d'un bruit indépendant, identiquement distribué et de loi continue $(\varepsilon_1, \cdots, \varepsilon_n)$ sa complexité est en $\mathcal{O}(nD \log(\frac{n}{D+1}))$ ce qui correspond à la complexité observée empiriquement. Nous obtenons ce résultat en faisant une connexion entre le nombre de ruptures candidates enregistrées par pDPA(1) et l'enveloppe convexe de la marche aléatoire $S_1 = Y_1$ et $S_{i+1} = S_i + Y_{i+1}$.

Mots-clés. Détection de ruptures multiples, élagage, marche aléatoire, enveloppe convexe

Abstract. The pDPA algorithm jointly estimates the location of K changes in a univariate signal by maximizing a likelihood. The pDPA has a quadratic worst-case complexity but empirically it is often log-linear. In this work we investigate the expected complexity of the pDPA algorithm ran for $K = 1$ changepoints: pDPA(1). We prove that for data $(Y_1, \cdots, Y_n)$ obtained summing a piecewise constant signal in $D + 1$ pieces $(\mu_1, \cdots, \mu_n)$ plus some independent and identically distributed noise with a continuous distribution $(\varepsilon_1, \cdots, \varepsilon_n)$ pDPA(1) complexity is in $\mathcal{O}(nD \log(\frac{n}{D+1}))$. This match the empirically observed complexity of pDPA(1). We obtain this result by making a connection between the set of candidate changepoints stored by pDPA and the convex hull of the random walk $S_1 = Y_1$ and $S_{i+1} = S_i + Y_{i+1}$.

Keywords. Multiple changepoint detection, pruning, random walk, convex hull

1 Introduction

In the last decade many approaches have been proposed for detecting changes in mean, see Truong et al. (2020) for a recent review of the area. One approach is to jointly estimate the location of all change-points by maximizing a penalized likelihood. Computationally, dynamic programming was proposed to solve this optimization. The problem can be solved either for a fixed number of changepoints K (Fisher, 1958; Bellman and Kotkin, 1962; Auger and Lawrence, 1989) or a fixed penalty per changepoint λ (Jackson et al., 2005).

Recently pruning rules were proposed to speed these algorithms. Killick et al. (2012) proposed the PELT algorithm, implementing an *inequality* pruning rules, which reduces time complexity from quadratic to linear in asymptotic regimes where the number of changepoints increases linearly with the number of data. A *functional* pruning rule was independently discovered by Johnson (2013) and Rigaill (2015) and implemented in the pDPA and FPOP algorithm (Maidstone et al., 2017).

Assuming both pruning rules can be applied, *functional* pruning always prunes more than *inequality* pruning, and empirically shows reduced time complexity in many situations. For example, it is shown in (Maidstone et al., 2017) that when we have data with no changes, the PELT algorithm has quadratic complexity, but functional pruning algorithms can have a log-linear complexity.

In this work, we study the number of candidate changepoints stored by the pDPA algorithm ran for $K = 1$ changepoints: pDPA(1). We prove that when we have data with no changes and an identically distributed and continuous distribution noise pDPA(1) store at most $2\sum_i^n \frac{1}{i}$ candidate changes. Using this bound we recover the observed log-linear complexity of the pDPA(1) algorithm. Considering data with D changes and again an identically distributed and continuous distribution noise we prove that pDPA(1) stores at most $2\frac{n}{D}\sum_1^{\frac{n}{D}} \frac{1}{i}$ changes.

2 Expected complexity of pDPA(1)

2.1 Data and definitions

We consider n datapoints with i in $\{1, ..., n\}$

$$Y_i = \mu_i + \varepsilon_i, \tag{1}$$

where ε_i are i.i.d with a continuous distribution and μ_i is a piecewise constant signal in $D + 1$ pieces with changes $\tau_0 = 0, \tau_1, \dots \tau_D, \tau_{D+1} = n$ and for i in $(\tau_{d-1}, \tau_d]$ $\mu_i = \theta_d$. We also define the length of the d-th segment l_d as $l_d = \tau_d - \tau_{d-1}$.

We define the random walk S as follow $S_1 = y_1$ and $S_{i+1} = S_i + y_{i+1}$.

We will often consider subsets of the data $i : j = \{i, \cdots, j\}$ and define $Y_{i:j}$ as $\{Y_i, \cdots, Y_j\}$. Similarly we define $S_{i:j}$ as the sequence $\{S_i, \cdots, S_j\}$.

2.2 Cost of a segmentation and pDPA with $K = 1$ change

We define the cost of a segmentation in two of $i : j = \{i, \cdots, j\}$ with a change at τ and means μ_1 and μ_2:

$$C_\tau(y_{i:j}, \mu_1, \mu_2) = \sum_{t=i}^{\tau}(y_t - \mu_1)^2 + \sum_{t=\tau+1}^{j}(y_t - \mu_2)^2.$$

By convention for j we take $C_j(y_{i:j}, \mu_1, \mu_2) = \sum_{t=i}^{j}(y_t - \mu_1)^2$.

The pDPA algorithm ran for 1 change (pDPA(1)) optimize this least square cost for $1 : n$:

$$\min_{\tau,\mu_1,\mu_2}\left\{\sum_{t=1}^{\tau}(y_t - \mu_1)^2 + \sum_{t=\tau+1}^{n}(y_t - \mu_2)^2\right\} = \min_{\tau,\mu_1,\mu_2}\left\{C_\tau(y_{1:n}, \mu_1, \mu_2)\right\}.$$

In details the pDPA does so by updating the following quantity

$$\widetilde{P}_{1:n}(\mu_2) = \min_{\tau,\mu_1}\left\{C_\tau(y_{1:n}, \mu_1, \mu_2)\right\}. \tag{2}$$

$\widetilde{P}_{1:n}(\mu_2)$ is a piecewise quadratic function in μ_2. A critical quantity of pDPA(1) is the number of pieces or intervals of $\widetilde{P}_{1:n}(\mu_2)$. This number is directly related to the number of changepoints τ actually reaching $\widetilde{P}_{1:n}(\mu_2)$ defined as follows.

$$\mathcal{I}_P(y_{1:n}) = \left\{\tau \mid \exists\mu_2, \forall\tau' \in (1:n)\setminus\{\tau\},\ \min_{\mu_1} C_{1:n,\tau}(\mu_1, \mu_2) < \min_{\mu_1} C_{1:n,\tau'}(\mu_1, \mu_2)\right\}. \tag{3}$$

The number of elements of this set: $\#\mathcal{I}_P(y_{1:n})$ control the complexity of the algorithm. Our goal is to bound its expectation. To this end, we consider a slight modification of the pDPA(1) problem.

2.3 A slightly modified problem

Rather than optimizing the value μ_1 of the first segment as in the pDPA(1) (see equation (2) we let it free and define:

$$\widetilde{M}(y_{i:j}, \mu_1, \mu_2) = \min_{\tau\ \in\ i+1:j}\{C_{i:j,\tau}(\mu_1, \mu_2)\}. \tag{4}$$

We can define the set of changes reaching $\widetilde{M}(y_{i:j}, \mu_1, \mu_2)$ as follows

$$\mathcal{I}_M(y_{i:j}) = \{\tau \mid \exists\mu_1, \mu_2, \forall\tau' \in (i+1:j)\setminus\{\tau\},\ C_{i:j,\tau}(\mu_1, \mu_2) < C_{i:j,\tau'}(\mu_1, \mu_2)\}. \tag{5}$$

It will make sense to consider the case $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$ separately. Therefore we also define the following sets:

$$\mathcal{I}_M^+(y_{i:j}) = \{\tau \mid \exists\mu_1 < \mu_2, \forall\tau' \in (i+1:j)\setminus\{\tau\}, \quad C_{i:j,\tau}(\mu_1, \mu_2) < C_{i:j,\tau'}(\mu_1, \mu_2)\},$$
$$\mathcal{I}_M^-(y_{i:j}) = \{\tau \mid \exists\mu_1 > \mu_2, \forall\tau' \in (i+1:j)\setminus\{\tau\}, \quad C_{i:j,\tau}(\mu_1, \mu_2) < C_{i:j,\tau'}(\mu_1, \mu_2)\}.$$

2.4 Some properties on $\mathcal{I}_M(y_{i:j})$

It is straightforward to see that the candidate stored by the pDPA(1), $\mathcal{I}_P(y_{1:n})$, are in $\mathcal{I}_M(y_{1:n})$ and we get the following lemma.

Lemma 2.1.

$$\mathcal{I}_P(y_{1:n}) \subseteq \mathcal{I}_M(y_{1:n}) = \mathcal{I}_M^+(y_{1:n}) \cap \mathcal{I}_M^-(y_{1:n}) \tag{6}$$

Furthermore using following identity

$$C_\tau(y_{i:j}, \mu_1, \mu_2) - C_{\tau'}(y_{i:j}, \mu_1, \mu_2) = (\mu_1 - \mu_2)\left(2\sum_{\tau+1}^{\tau'} y_i - \mu_1 - \mu_2\right) \tag{7}$$

we get the following inclusion lemma.

Lemma 2.2. *For $i \leq j \leq k$*

$$\mathcal{I}_M^+(y_{i:k}) \subset \mathcal{I}_M^+(y_{i:j}) \cap \mathcal{I}_M^+(y_{j+1:k}) \tag{8}$$

$$\mathcal{I}_M^-(y_{i:k}) \subset \mathcal{I}_M^-(y_{i:j}) \cap \mathcal{I}_M^-(y_{j+1:k}) \tag{9}$$

Proof. Consider any τ in $(i+1:j) \cap \mathcal{I}_M^+(y_{i:k})$, using equation (7) we get that

$$\exists \mu_1, \mu_2, \forall \tau' \in (i+1:j) \setminus \{\tau\}, \quad C_{i:j,\tau}(\mu_1, \mu_2) < C_{i:j,\tau'}(\mu_1, \mu_2),$$

and therefore τ is also in $\mathcal{I}_M^+(y_{i:j})$. □

The following lemma relates these sets to the convex hull of $S_{i:j}$.

Lemma 2.3. *The set of τ in $\mathcal{I}_M^+(y_{i:j})$ are the extreme points of the largest convex minorant of the sequence $S_{i:j}$. By symmetry, the set of τ in $\mathcal{I}_M^-(y_{i:j})$ are the extreme points of the smallest concave majorant of S_i*

Proof. Using equation (7), we get that if τ' is in $\mathcal{I}_M^+(y_{i:j})$ it must be that for any τ and τ'' $\bar{y}_{\tau+1:\tau'} < \bar{y}_{\tau'+1:\tau}$. This is equivalent to for all τ and τ'': $\frac{S_{\tau'} - S_\tau}{\tau' - \tau} < \frac{S_{\tau''} - S_{\tau'}}{\tau'' - \tau'}$. □

2.5 Expectation on the number of changepoints stored by pDPA(1)

Here is our my main lemma bounding the number of candidate changes stored by pDPA(1).

Lemma 2.4. *Considering data defined as in equation* (1) *We get*

$$E(\#\mathcal{I}_P^+(y_{1:n})) \leq E(\#\mathcal{I}_M^+(y_{1:n})) = \sum_{d=1}^{D+1} \sum_{1}^{l_d} 1/t \leq (D+1) \sum_{1}^{n/(D+1)} 1/t$$

Proof. The first inequality is given by lemma 2.1. We then use lemma 2.2 to get that

$$\mathcal{I}_M^+(y_{1:n}) \subseteq \bigcap_{k \in 1:K} \mathcal{I}_M^+(y_{\tau_{k-1}:\tau_k})$$

and apply results of Andersen (1955); Abramson (2012) on the number of extreme points of random-walks. □

2.6 Conclusion

From this we recover that the pDPA(1) will store at most $\mathcal{O}((D+1)\log(\frac{n}{D+1}))$ change-points and we recover the overall $\mathcal{O}(n(D+1)\log(\frac{n}{D+1}))$ complexity. In the rest of the presentation, we will demonstrate that our bound is tight in the sense there are some data defined as in equation (1) for which this complexity is achieved.

We also empirically study the complexity of pDPA(1) as a function of n and D for various types of signals. Below we describe three such signals with an i.i.d Gaussian noise of variance 1.

no-change-case We considered $\mu_i = 0$ for all i and vary n in $\{2^{12}, \cdots, 2^i, \cdots 2^{24}\}$. We observed in figure 1-(left) that the average runtime of pDPA(1) is quasi-linear in n.

stair-case We considered stair-case signals of size $n = 10^5$ and D segments with a deterministic mean $\mu_i = \lceil Di/n \rceil$. We obvserved in figure 1-(middle) that the average runtime is almost linear in D.

random-case We also considered signals of size $n = 10^5$ with D segments of equal length with the mean of each segment sampled independently from a Gaussian distribution of variance 100. We observed in figure 1-(right) a reduced average runtime and limited linear dependency in D compared to the stair-case showing that our complexity is, at least in some cases, pessimistic.

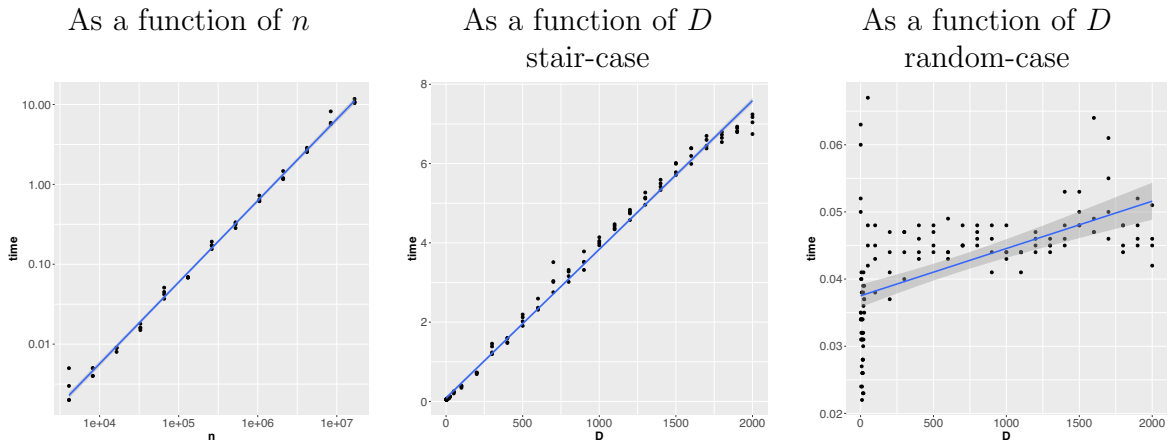


Figure 1: Empirical average runtime of pDPA(1) (left) as a function of n for signals without changes and n in $[2^{12}, 2^{24}]$ (middle) as a function of D in $[1, 2000]$ for a stair-case signal (right) as a function of D for signal with D equal-length segments with randomly picked means and D in $[1, 2000]$. In all three panels, dots represent the runtime (on a laptop with four 2.10 GHz cores) for one simulated signal and the blue line the fit using a simple linear model.

References

Joshua Simon Abramson. *Some Minorants and Majorants of Random Walks and Lévy Processes*. PhD thesis, UC Berkeley, 2012.

Erik Sparre Andersen. On the fluctuations of sums of random variables ii. *Mathematica Scandinavica*, pages 195–223, 1955.

Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.

Richard Bellman and Bella Kotkin. On the approximation of curves by line segments using dynamic programming. ii. Technical report, RAND CORP SANTA MONICA CALIF, 1962.

Walter D Fisher. On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798, 1958.

Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

Nicholas A. Johnson. A Dynamic Programming Algorithm for the Fused Lasso and L0-Segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

Guillem Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, 156(4):180–205, 2015.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Krigeage Monte Carlo : prise en compte de données localisées aux mêmes points.

Luc Rongiéras [1,2] & Emilie Chautru [2]

[1] *CCR, 157 Boulevard Haussmann 75008 Paris, lrongieras-prestataire@ccr.fr*
[2] *MINES ParisTech, Université PSL, Centre de Géosciences, 35 rue Saint Honoré 77300 Fontainebleau, emilie.chautru@mines-paristech.fr*

Résumé. L'estimation d'une fonction aléatoire en un point de l'espace peut être donnée par les méthodes de Krigeage, qui visent à construire une interpolation linéaire des données. En pratique, il est possible que la géolocalisation comporte des erreurs. On s'intéresse ici au cas particulier où plusieurs données sont géolocalisées à tort au même point. Dans ce cas, les méthodes classiques de Krigeage ne sont plus applicables. Nous proposons ici de les adapter à ce contexte particulier en nous inspirant des approches Monte-Carlo. En considérant les géolocalisations aléatoires, l'approche développée consiste à en simuler plusieurs réalisations. Pour chacune d'entre elles, les poids de Krigeage sont calculés, puis les résultats sont moyennés. Une étude sur données simulées illustre l'efficacité de la méthode pour estimer la structure de dépendance spatiale. Elle est enfin appliquée à des données assurantielles.

Mots-clés. Géostatistique, Monte-Carlo, Krigeage, Simulation, Erreur de localisation.

Abstract. The estimation of a random function at a spatial location can be given by Kriging methods, which aim to build a linear interpolation of the available data. In practice, the observations may suffer from a misplacement of their true location. A deep interest is taken here in the case where they are mistakenly assigned to a single point in space. In that case, classic Kriging methods are not applicable. We propose to adapt them to this particular context, using the scheme of Monte Carlo approaches. Considering all locations as random, it consists in first simulating several of their realizations, then computing the Kriging weights for each of these simulations and finally averaging them. A study with simulated data illustrates the efficiency of the estimation of the variogram based on this method. Then, it is applied to insurance data.

Keywords. Geostatistics, Monte-Carlo, Kriging, Simulation, Location error.

1 Introduction

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. L'objet d'intérêt est une fonction aléatoire Z définie sur un espace d'indexes $D \subset \mathbb{R}^2$, supposée stationnaire du second ordre.

1.1 Cadre standard

Supposons que Z est observée en $N \in \mathbb{N}^*$ points de l'espace $\mathbf{x}_N = (x_1, ..., x_N) \in D^N$; on note $\mathbf{Z}_N = (Z(x_1), ..., Z(x_N))$ le vecteur de données. La méthode du Krigeage ordinaire propose d'estimer (interpoler) la valeur de Z en un point quelconque $x_0 \in D$ par une combinaison affine des données. Précisément, on cherche un vecteur de poids $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$ et une constante $\lambda_0 \in \mathbb{R}$ tels que l'estimateur $\widehat{Z}(x_0) = \lambda_0 + \lambda^\intercal \mathbf{Z}_N$ est sans biais, *i.e.* $\mathbb{E}[\widehat{Z}(x_0) - Z(x_0)] = 0$, et d'erreur quadratique moyenne $\mathbb{E}[(\widehat{Z}(x_0) - Z(x_0))^2]$ minimale.

Lorsque tous les points d'observation $x_1, \ldots, x_N$ sont distincts, la solution à ce problème d'optimisation est unique (Chilès et Delfiner, 2012, section 3.4.1 pp.163-164) et s'exprime en fonction du variogramme (ou semivariogramme) :

$$\gamma : \begin{array}{ccc} D \times D & \longrightarrow & \mathbb{R}_+ \\ (x, x') & \longmapsto & \frac{1}{2}\mathbb{V}[Z(x) - Z(x')]. \end{array}$$

Plus précisément, $\lambda_0 = 0$ et le vecteur λ optimal dépend des valeurs $\{\gamma(x_i, x_j)\}_{0 \leq i,j \leq N}$. Ainsi, $\lambda = \lambda(\mathbf{x}_N, x_0)$ est fonction des positions initiales $\mathbf{x}_N$ et de la position cible x_0, dont la connaissance est cruciale.

1.2 Cadre d'étude

Lorsque les localisations $\mathbf{x}_N$ et x_0 sont erronées, le Krigeage ordinaire ne peut plus être appliqué directement. Des méthodes prenant en compte une potentielle erreur de localisation ont ainsi été développées, notamment dans Chilès (1976) puis Gabrosek (2002), ou encore de manière plus extensive dans Gabrosek (1999). Cependant, à notre connaissance, aucune ne considère le cas où plusieurs positions sont assignées à un même point de référence. Certaines méthodes, telles que l'utilisation d'un modèle gaussien discret, s'approchent de cette configuration spécifiques. Néanmoins, une différence reste notable, que ce soit dans les hypothèses initiales ou dans les objectifs.

Or ce cas de figure peut apparaître, en particulier, dans des données d'assurance immobilière. Ces dernières peuvent en effet être divisées en $n \in [\![1, N]\!]$ groupes, auxquels sont affectées des localisations de référence $\overline{\mathbf{x}}_n = (\overline{x}_1, ..., \overline{x}_n)$. Par exemple, des bâtiments peuvent être référencés au centre de leur quartier.

Dans ce cas, la fonction de référencement $\tau : [\![1, N]\!] \to [\![1, n]\!]$ est généralement connue. Pour un indice $i \in [\![1, N]\!]$, $\tau(i)$ correspond à l'indice de la position $\overline{x}_{\tau(i)}$ à laquelle la position réelle x_i est référencée. Pour chaque $j \in [\![1, n]\!]$, on a donc $m_j := \mathrm{Card}(\tau^{-1}(\{j\}))$ positions réelles référencées en $\overline{x}_j$. Comme illustré en figure 1, on peut alors définir une série de voisinages $A_1, ..., A_n \subset D$ deux à deux disjoints autour des positions de référence telle que

$$\forall i \in [\![1, N]\!], \quad \forall j \in [\![1, N]\!], \qquad \tau(i) = j \iff x_i \in A_j.$$

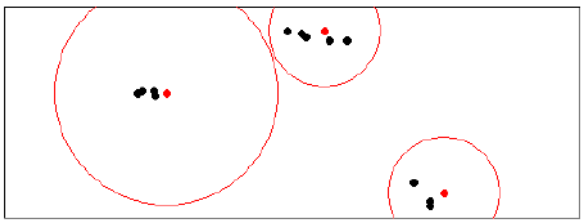


FIGURE 1 – Exemple schématique d'une configuration possible de regroupement de localisations : les points noirs représentent les positions réelles, les points rouges les positions de référence et les cercles rouges des voisinages possibles.

Notre objectif dans ce cadre est le suivant : pour toute localisation de référence $\overline{x}_0$ et son voisinage A_0, on cherche à estimer $Z(x_0)$ pour un point x_0 choisi aléatoirement dans A_0.

2 Méthode

Les positions réelles $\mathbf{x}_N, x_0$ étant inconnues, on les suppose aléatoires, de lois dont les supports correspondent à leurs voisinages $A_1, ..., A_n, A_0$ respectifs. En notant $\lambda : D^N \times D \to \mathbb{R}^N$ la fonction de poids optimale obtenue dans le cadre classique, nous proposons de chercher un estimateur $\widehat{\lambda}$ de $\lambda(\mathbf{x}_N, x_0)$ indépendant des positions réelles. On souhaite qu'il soit sans biais et minimise la quantité

$$\mathbb{E}\left[\left\|\lambda(\mathbf{x}_N, x_0) - \widehat{\lambda}\right\|^2\right],$$

ce qui est obtenu pour $\widehat{\lambda} = \mathbb{E}[\lambda(\mathbf{x}_N, x_0)]$ sous réserve que ces quantités existent. Ce sera le cas en particulier sous les conditions énoncées dans le théorème 2.1 suivant.

Théorème 2.1 *Soit Z une fonction aléatoire stationnaire isotrope du second ordre de variogramme γ dérivable sur $\mathbb{R}_+^*$ possédant un effet de pépite $\gamma_0 > 0$. Supposons qu'elle est observée en $N \in \mathbb{N}^*$ localisations aléatoires, admettant une densité bornée. Alors tous les moments des N poids de Krigeage existent.*

Néanmoins, le calcul de $\mathbb{E}[\lambda(\mathbf{x}_n, x_0)]$ peut s'avérer fastidieux, si bien qu'on lui préfèrera son équivalent empirique par souci de simplicité.

La méthode de Krigeage Monte-Carlo suit finalement le pseudo-code suivant :

1. Simuler $p \in \mathbb{N}^*$ jeux de localisations $(\mathbf{x}_N^{(1)}, x_0^{(1)}), ..., (\mathbf{x}_N^{(p)}, x_0^{(p)})$
2. Pour chaque simulation $q \in [\![1, p]\!]$, calculer les poids de Krigeage $\lambda^{(q)} := \lambda(\mathbf{x}_N^{(q)}, x_0^{(q)})$

3. Calculer la moyenne $\widehat{\lambda} = \frac{1}{p}\sum_{q=1}^{p} \lambda^{(q)}$ et en déduire l'estimateur $\widehat{Z}(x_0) = \widehat{\lambda}^{\mathsf{T}}\mathbf{Z}_N$.

Le théorème 2.1 et la loi des grands nombres assurent la convergence presque-sûre de l'estimateur empirique vers $\mathbb{E}[\lambda(\mathbf{x}_n, x_0)]^{\mathsf{T}}\mathbf{Z}_N$.

3 Application sur données simulées : pertinence de l'estimation du variogramme

Lorsque l'on souhaite proposer une estimation via Krigeage de $Z(x_0)$, il est nécessaire de connaître le variogramme γ. En pratique, le variogramme théorique n'est pas connu et est remplacé par une estimation. Dans le cadre du Krigeage Monte-Carlo, l'estimation du variogramme est perturbée par l'ignorance des localisations réelles.

Plusieurs variogrammes peuvent être estimés et choisis pour représenter la structure spatiale. Lorsque les données sont connues, on estime un variogramme $\widehat{\gamma}_{\text{réel}}$. Dans le cadre du Krigeage Monte-Carlo, on peut choisir pour la simulation de positions q le variogramme calculé à partir des localisations simulées $\widehat{\gamma}^{(q)}$. Il est également possible de calculer la moyenne empirique de Z aux localisations de référence $\overline{x}_j$ avec $m_j > 1$, et d'utiliser les résultats pour construire un variogramme. Cette méthode nous donne le variogramme agrégé $\widehat{\gamma}_{\text{agr}}$. Ces techniques sont comparées en figure 2 dans le cadre de données simulées d'un champ gaussien avec un variogramme théorique exponentiel de portée 0.6. On remarquera que Z étant faiblement stationnaire, son variogramme ne dépend que de la distance euclidienne entre les couples de points qu'il prend en entrée ; on peut alors le représenter comme une fonction d'une seule variable.

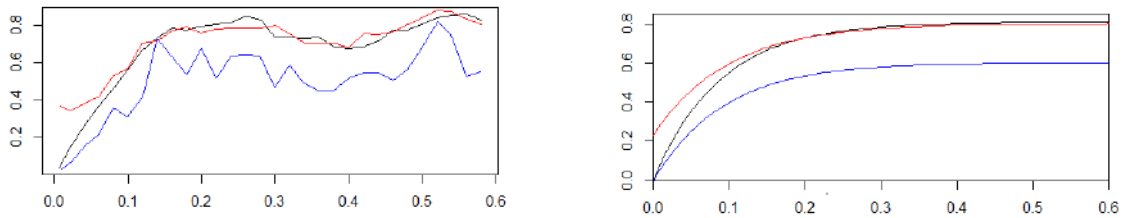


FIGURE 2 – Comparaison des estimateurs des variogrammes $\widehat{\gamma}_{\text{réel}}$ (en noir), $\widehat{\gamma}_{\text{agr}}$ (en bleu) et $\widehat{\gamma}^{(q)}$ (en rouge) en fonction de la distance entre les points de l'espace, calculés de manière non-paramétrique (à gauche) puis paramétrique (à droite).

Que l'estimation soit paramétrique ou non, on retrouve les mêmes tendances :
— $\widehat{\gamma}_{\text{agr}}$ garde l'effet de pépite mais sous-estime $\widehat{\gamma}_{\text{réel}}$,
— $\widehat{\gamma}^{(q)}$ surestime $\widehat{\gamma}_{\text{réel}}$ pour les petits écarts (notamment l'effet de pépite) mais reste proche aux alentours de la portée.

4 Application sur données réelles : estimation de la surface des appartements dans la ville de Lyon

L'application de la méthode d'estimation par Krigeage Monte-Carlo peut s'effectuer dans différents cadres dès lors que plusieurs localisations sont référées au même endroit. Cela peut-être dû à une incertitude sur des capteurs lors de la collecte de données pour une variable physique, la perte d'information concernant une des dimensions de l'espace d'index. D'autres configurations peuvent être imaginées.

On se place maintenant dans un cadre d'étude assurantiel avec des données réelles. L'espace d'index D considéré est une projection locale (Lambert93) de la ville de Lyon. La fonction aléatoire Z est la surface des appartements.

Certaines études exploratoires préliminaires nous amènent à supposer Z stationnaire du second ordre et isotrope. On choisit un échantillon de $N = 7000$ données pour lesquelles on connaît la surface, et un échantillon de 1000 données pour lesquelles on cherche à l'estimer.

L'estimation est effectuée en considérant $p = 30$ simulations de localisations dans les disques de rayon $\rho = 4$m autour de leurs positions de référence. Le variogramme utilisé $\widehat{\gamma}_{\text{agr}}$ est construit en prenant la moyenne des surfaces aux mêmes localisations.

On retient alors deux indicateurs d'intérêt :
- L'écart-type $\widehat{\sigma}_\mu$ des 1000 erreurs d'estimation par Krigeage Monte-Carlo de l'échantillon test. Il représente la variabilité sur D de l'estimation par Krigeage Monte Carlo.
- La moyenne $\widehat{\mu}_\sigma$ des 1000 écarts-types de l'erreur d'estimation, calculés sur les 30 simulations pour chaque localisation de référence. Il représente la moyenne de la variabilité d'estimation induite par les simulations des localisations sur D.

En parallèle, on construit $\widehat{\sigma}_Z$ qui est l'écart-type des erreurs d'estimation obtenues en choisissant la moyenne de l'ensemble des 7000 données d'apprentissage comme estimateur constant de la surface des appartements. On obtient les résultats fournis en Table 1.

$\widehat{\mu}_\sigma$	$\widehat{\sigma}_\mu$	$\widehat{\sigma}_Z$
1.26	27.49	29.80

TABLE 1 – Résultats du Krigeage Monte Carlo appliqué au calcul de la surface des appartments de la ville de Lyon.

La faible valeur de $\widehat{\mu}_\sigma$ indique que l'estimation de la surface des appartements pour chaque simulation de localisation change peu sur D. De ce fait, on peut choisir un nombre

simulations p relativement faible afin d'accélérer le processus de calcul.

La quantité $\widehat{\sigma}_\mu$ est légèrement plus faible que $\widehat{\sigma}_Z$, ce qui suggère que l'estimateur par Krigeage Monte-Carlo est meilleur que celui, naïf et constant, reposant sur la moyenne des données d'apprentissage, mais garde une valeur ajoutée limitée. Cela couplé au fait que $\widehat{\mu}_\sigma$ est faible nous laisse penser que la structure spatiale est difficile à exploiter dans ce jeu de données, sans doute en raison de la présence d'un fort effet de pépite inhérent.

5 Conclusion et perspectives

Les résultats obtenus via la méthode de Krigeage Monte Carlo sont encourageants. Elle dépend toutefois de certains paramètres, en particulier du variogramme théorique, de la zone et du nombre de simulations, ou encore du nombre de localisations regroupées, dont l'influence exacte reste à étudier précisément. En outre, des améliorations peuvent être apportées dans l'estimation du variogramme : un mélange entre les estimateurs agrégé $\widehat{\gamma}_{\text{agr}}$ et aléatoire $\widehat{\gamma}^{(q)}$ pourrait peut-être nous permettre de mieux en estimer le comportement à l'origine.

Enfin, cette méthode permet d'étendre le Krigeage Ordinaire à un cadre non considéré initialement. Il serait intéressant de regarder s'il est possible d'adopter un schéma similaire pour des méthodes nécessitant des hypothèses plus souples, comme le Krigeage Universel ou Krigeage disjonctif. L'extension au cas non-stationnaire pourrait également permettre d'appliquer le Krigeage Monte Carlo à une plus grande catégorie de fonctions aléatoires, ie : modèles de mélange spatial, fonctions aléatoires IRF-k, ...

Bibliographie

Chilès, J. P. et Delfiner, P. (2012), *Geostatistics, Modelling Spatial Uncertainty*, Wiley.

Gabrozek, J. et Cressie, N. (2002) , *The Effect on Attribute Prediction of Location Uncertainty in Spatial Data*, Wiley.

Chiles, J. P. (1976). *How to adapt kriging to non-classical problems : Three case studies.* In *Advanced Geostatistics in the Mining Industry* (M. Guarascio, M. David and C. Huijbregts, eds.) 69–89. Reidel, Dordrecht.

Gabrosek, J. (1999). *The effect of locational uncertainty in geostatistics.* Ph.D. dissertation, Dept. Statistics, Iowa State Univ., Ames.

Estimation of the average treatment effect of the restricted survival time with censored data and missing values

Paul Roussel [1,2], Julie Josse [1] Sebastien Bernard [2]

[1] *Inria Sophia Antipolis,* [2] *Sanofi-Aventis RD*

Résumé Nous nous intéressons à l'estimation de l'effet moyen du traitement sur le temps de survie limité de données censurées à droite, à partir d'une étude observationnelle. Nous examinons et comparons différentes méthodes d'estimation causale telle que les méthodes de probabilité inverse, la méthode de la formule-g, ainsi que des méthodes multiples robustes. À partir d'une étude de simulation approfondie, nous soulignons leurs différences avec des méthodes populaires telles que Kaplan-Meier, Kaplan-Meier ajusté. Nous proposons différentes approches pour gérer les valeurs manquantes dans les covariables. Notre étude est motivée par une étude médicale sur l'effet de la transfusion sur la mortalité à un an pour des patients en soins intensifs.

Mots-clés

Inférence causale, Analyse de survie, Données censurées, Données manquantes, robuste, Données de santé, soins intensifs

Abstract. We are interested in the estimation of the average treatment effect of the restricted survival time on right-censored data from an observational study. We review and compare different causal estimation methods such as inverse probability methods, the g-formula method, as well as multiply robust methods using an extensive simulation study and highlight their differences with popular methods such as Kaplan-Meier or the adjusted Kaplan-Meier. We propose different approaches to deal with missing values in covariates. Our study is motivated by a medical study on the effect of transfusion on one-year mortality for patients in intensive care.

Keywords.

Causal inference, Survival analysis, Censored data, Missing data, Multiply robust, health data, ICU

1 Introduction

Our study is motivated by a medical application where the aim is to estimate the effect of transfusion on the survival time (1-year mortality) for patients in intensive care unit (ICU). The data can be defined as right-censored survival data, also known as time to event data, where patients are followed over a certain period only and there is a drop in follow-up, leading to censorship.

The gold standard for estimating the effect of a treatment on an outcome in randomized clinical trials (RCTs) in which the treatment is given independently to the patient's features. However, in some domains such as in intensive care, it may not be possible to conduct a RCT due to ethical or practical reasons. We, therefore, rely on observational data, where the allocation of the treatment is not under control, and we need to take into account confounding biases in the estimation. In observational medical studies, as in many studies, missing values in the covariates are ubiquitous and occur for many reasons (measures that have not been recorded, the state of the patient was such that it was not possible to make the measurement, etc..). Naive methods such as complete case can lead to the deletion of almost all data while imputation methods can distort the distributions. Hence, we encounter three types of possible biases in the estimation: due to the allocation of the treatment, due to the missing data, and due to the censoring process and need to design appropriate strategies to estimate the effect of the treatment?

In the presentation, we review and detail three classes of consistent estimators of the restricted mean survival time. We highlight their differences with classical methods such as Kaplan-Meier or the adjusted Kaplan-Meier that do not have a causal interpretation on observational data and compare the estimators with an extensive simulation study. Finally, we illustrate their different behaviors on the medical data that underline the importance of discussing their underlying assumptions.

2 Estimators of the restricted mean survival time

2.1 Setting

We observe an independent and identically distributed sample $(O_i, \cdots, O_n)$ with $O_i = \left(X_i, A_i, \Delta_i, \tilde{T}_i\right)$ where X is a p dimensional covariate, $A \in \{0, 1\}$ is the binary treatment, $\tilde{T}$ is the right-censored event time, i.e., $\tilde{T} = T \wedge C$ where T is the event time, C the censoring time, Δ is the event type for which we assume that $\{\Delta = 1\}$ means that the event of interest occurred and $\{\Delta = 0\}$ indicates uncensored observation. We consider the potential outcome framework (rubin) where we define the potential outcome (T_1, T_0) as the event times that would have been observed had we assigned treatment respectively control to each observation.

As we focus on the observational study setting, we make the following four assumptions on the failure time T for identifiability of the average treatment effect.

Assumption 1. (Consistency) $T = A \cdot T(1) + (1 - A) \cdot T(0)$ almost surely.

Assumption 2. (Unconfoundedness) $\{T(0), T(1)\} \perp W \mid X$.

Assumption 3. (Positivity) $1 > p(A = a \mid X = x) > 0$ & $p(C \geq t \mid X = x, A = a) > 0$

Assumption 4. (Conditionally independent censoring) $T \perp C \mid X, A$.

Given τ a threshold time, we want to estimate the treatment effect defined as:

$$\theta = E[T(1) \wedge \tau - T(0) \wedge \tau].$$

It is generally called the restricted mean survival time (RMST), the expected survival time restricted to a time τ. This parameter [3, 6, 7] has a more straightforward interpretation compare to the hazard ratio that is widely used in practice. Indeed for $\tau = 365$ days, for instance, a 10-day improvement in RMST due to treatment means 10 more days alive on average during one year; this may be more directly interpretable than a hazard ratio.

2.2 Estimators

We propose an in-depth review that compares different estimators of restrictive mean survival time and discusses their underlying assumptions. We focus on three classes of estimators as in the uncensored causal inference framework: G-formula, inverse probability weighting (IPW), and augmented inverse probability weighting methods. These estimators rely on different combination of what is called nuisance models, namely an outcome model $F(a, x) = E[T \wedge \tau | X = x, A = a]$ computed with a model of the survival function, a treatment model $e(x) = E[A = 1 \mid X = x]$ and a censoring model $S^C(t, x, a) = p(C \geq t \mid X = x, A = a)$. The models for the conditional survival function and the censoring function can be estimated either using non-parametric approaches using Kaplan-Meier and Nelson-Aalen method, semi-parametric approaches using cox regression but also fully parametric approaches with accelerated Failure Time models. Recently have been developed survival forest regression [4]. These different models are based on different assumptions and may not be consistent estimators of the nuisances parameters.

The first class of estimators, G-formula estimator, is solely based on the outcome model:

$$\widehat{\theta}_{\mathrm{G-formula}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{F}\left(1, X_i\right) - \hat{F}_{1n}\left(0, X_i\right) \right).$$

The second class of estimators uses inverse probability of censoring or/and treatment weights (IPW). We detail one estimator in this class, that is obtained from the inverse probability of treatment weights estimator in the case of uncensored data:

$$\widehat{\theta}_{\mathrm{IPW}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \frac{1\left\{\tilde{T}_i \leq \tau, \tilde{\Delta}_i \neq 0\right\}}{\hat{S^C}\left(\tilde{T}_i \mid A_i, X_i\right)} T_i \wedge \tau(\tau) \left(\frac{A_i}{\hat{e}_n\left(X_i\right)} - \frac{1 - A_i}{1 - \hat{e}_n\left(X_i\right)} \right).$$

Both classes lead to consistent estimators of the RMST when the nuisance terms are consistently estimated.

The third class of method rely on augmented inverse probability methods and enjoy better statistical properties: they are known to be multiple robust [5, 7, 10, 11] meaning for instance that the RMST estimator will be consistent if either only the model of the outcome or the models of the censoring and the treatment are consistent. A recent promising estimator in this class is the survival causal forest [11].

3 Missing data

We extend the previous estimators so that they can handle missing covariate in order to have a complete methodology for users to estimate the effect of the treatment from incomplete observational data. Although there are many methods available in the literature to deal with missing values, there are only a few strategies available in the context of treatment effect estimation for observational data (under unconfoundedness) [9]. Note that these approaches are dedicated to uncensored data and have not been considered in the framework of RMST estimation.

Under the assumption of random censoring and classical assumptions on the missing values mechanism (MAR, MCAR), we consider a multiple imputation method in which several imputed data-sets are generated and the RMST estimators are applied on each of the imputed data-sets. Then we combine the results using the Rubin's Rules[1].

And as an alternative, and under modified identifiability assumptions and no assumption on the missing value mechanism, when using survival forest and regression with random forest, we consider the missing incorporated in attributes (MIA) imputation method [8]. Indeed random forests [2] can handle semi-continuous variables therefore allowing for missing values in the data.

4 Simulation study

We compare the results of the different estimators with different data-generating processes for time-to-event outcomes and missing information in the covariates. We change the data generating process to challenge the different assumptions concerning the unconfoundedness, the positivity assumptions, and the random censoring. We also numerically compare the different estimators with the standard approach used in clinical trials that do not necessarily identify the estimand given the assumption in observational study: Kaplan-Meier method or adjusted Kaplan-Meier methods.

For the simulations we generate $O_i = (X_i, A_i, \Delta = 1(T_i \leq \tau), \tilde{T} = T \wedge \tau)$. We generate a baseline covariate vectors, $X \in R^n$ as multivariate normal with mean $\mu \in R^n$, and variance $\Sigma \in R^{n \times n}$, then we simulate the treatment indicator A using a logistic regression given the covariates. Finally, we simulate the censoring probability and the survival probability from a proportional hazard models.

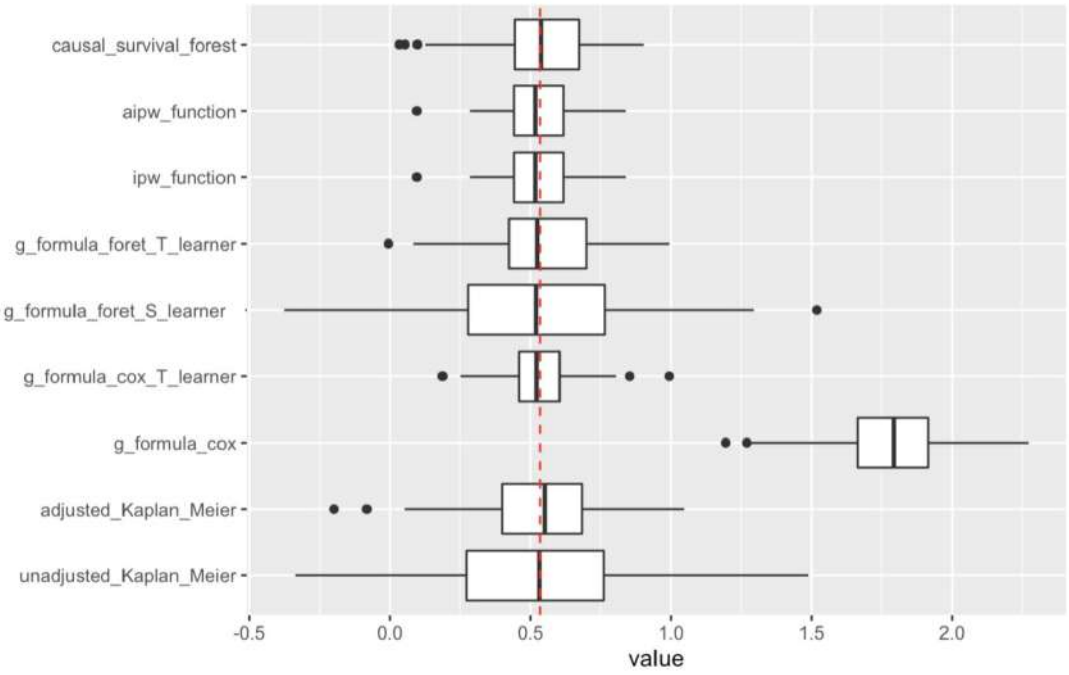


Figure 1: Boxplot of the results of 100 simulations for different estimators of the RMST (the true value of the RMST is the dotted line) . The treatment and the censoring is here taken independent of the survival time.

5 Application

The study is motivated by a dataset of patients joining ICU and include in the study. We have 1557 patients followed from their entry in ICU to one year after discharge. We observe a drop in follow-up for some patients after discharge. The patients are described by their clinical signs and biology at inclusion in the study, and then regularly for their all stay in ICU, by their medication before inclusion, by their chronic treatment, and by their comorbidities and the diagnosis at inclusion. The covariates can be quantitative and categorical and the percentage of missing values varies from 45 % for some clinical signs to 0 %. During their stay in ICU after inclusion, 663 patients received a transfusion and 894 didn't receive a transfusion. We are interested in estimating the effect of this transfusion on the 1-year mortality in terms of RMST. The allocation of the treatment was decided from a set of observed covariates, some covariates are responsible for the allocation of the treatment and also related to mortality. Therefore the allocation of the treatment is not under control, and we need to take into account confounding biases in the estimation. We use the estimators described above to estimate the effect and most methods point to a negative effect of the transfusion on the 1-year mortality.

References

[1] Donald B. Rubin, ed. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., June 1987. DOI: 10.1002/9780470316696. URL: https://doi.org/10.1002/9780470316696.

[2] Leo Breiman. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324. URL: https://doi.org/10.1023/a:1010933404324.

[3] Pei-Yun Chen and Anastasios A. Tsiatis. "Causal Inference on the Difference of the Restricted Mean Lifetime Between Two Groups". In: *Biometrics* 57.4 (Dec. 2001), pp. 1030–1038. DOI: 10.1111/j.0006-341x.2001.01030.x. URL: https://doi.org/10.1111/j.0006-341x.2001.01030.x.

[4] Hemant Ishwaran et al. "Random survival forests". In: (2008). DOI: 10.1214/08-AOAS169. eprint: arXiv:0811.1645.

[5] K. L. Moore and M. J. van der Laan. "Increasing power in randomized trials with right censored outcomes through covariate adjustment". In: *J Biopharm Stat* 19.6 (Nov. 2009), pp. 1099–1131.

[6] D. H. Kim, H. Uno, and L. J. Wei. "Restricted Mean Survival Time as a Measure to Interpret Clinical Trial Results". In: *JAMA Cardiol* 2.11 (Nov. 2017), pp. 1179–1180.

[7] I. Díaz et al. "Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards". In: *Lifetime Data Anal* 25.3 (July 2019), pp. 439–468.

[8] Julie Josse et al. *On the consistency of supervised learning with missing values*. 2019. eprint: arXiv:1902.06931.

[9] Imke Mayer et al. *Doubly robust treatment effect estimation with missing attributes*. 2019. eprint: arXiv:1910.10624.

[10] Brice Maxime Hugues Ozenne et al. "On the estimation of average treatment effects with right-censored time to event outcome and competing risks". In: (2019). DOI: 10.1002/bimj.201800298. eprint: arXiv:1907.12912.

[11] Yifan Cui et al. *Estimating heterogeneous treatment effects with right-censored data via causal survival forests*. 2020. eprint: arXiv:2001.09887.

Prediction intervals on individual electrical load curves using Bayesian neural networks

Honorine Royer [1] & Anne Philippe [2] & Philippe Charpentier [3] & Laurent Bozzi [4]

[1] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, honorine.royer@edf.fr*
[2] *Laboratoire de Mathématiques Jean Leray - Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, anne.philippe@univ-nantes.fr*
[3] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, philippe.charpentier@edf.fr*
[4] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, laurent.bozzi@edf.fr*

Résumé. Nous utilisons des méthodes d'apprentissage statistique afin de construire des intervalles de prévision sur des courbes de charge électrique individuelles à partir de variables clients. Plus précisément, nous utilisons un auto-encodeur pour faire de la réduction de dimension sur les courbes. Nous appliquons des réseaux de neurones bayésiens pour estimer la fonction de régression expliquant la couche latente de l'auto-encodeur à l'aide des informations clients. Nous adaptons ces méthodes pour un cas d'usage EDF sur la prévision de consommation d'électricité de clients non résidentiels durant les heures d'ensoleillement, pour du dimensionnement de panneaux photovoltaïques.

Mots-clés. Apprentissage profond, autoencodeurs, production solaire, réduction de dimension, statistique bayésienne

Abstract. We use machine learning methods to build prediction intervals on individual electrical load curves using customers information. Specifically, an autoencoder is used to reduce the dimension on the curves. Bayesian neural networks are used to estimate the regression function between the latent space of the autoencoder and the customers' information. Using the posterior predictive distribution we describe two possibilities for obtaining prediction intervals. We adapt the methods to a use-case from EDF concerning the prediction of electricity consumption of non-residential customers during hours of sunlight, for sizing of the customers' potential photo-voltaic installations.

Keywords. Autoencoders, Bayesian analysis, deep learning, dimensionality reduction, solar output generation...

1 Introduction

The opening of the French electricity market allows utility companies like EDF to develop innovative services for their non residential customers such as retail stores or farming businesses.

We present an industrial use case designed to build prediction intervals on annual load curves at half hourly period for non residential customers while emphasizing on

the prediction of hours of sunlight for photovoltaic sizing of potential installations. The prediction intervals are obtained for each half-hour of the curves.

Our dataset contains two categories of customers segmented according to their contract power, namely $C2$ and $C4$ consumers. The proportion of both categories is the following: 93% of $C2$ consumers and 7% of $C4$ consumers. We wish to predict the $C4$ customers' load curves using their billing information and their similarities with the $C2$ customers, as the $C4$ subset is too small. Despite sharing some resemblance, the $C2$ and $C4$ have different off-peak hours span, our models are designed to acknowledge those specificities. Precisely, we apply fine tuning to our models, a transfer learning technique where the models are first pre-trained on the $C2$ subset, and then the learned parameters on $C2$ are used as initialization for training of the model on $C4$. For training, we consider a subset containing 80% of the $C4$ consumers and the remaining $C4$ constitutes the testing subset, used to test and evaluate the methods. The $C2$ customers have bigger consumption volumes than the $C4$ yet we are interested in the load profile rather than the volume, thus all the load curves are normalized.

Following the industrial stake to correctly predict hours of sunlight, we use a weighted loss function where the weights are built from solar power generation data collected from several solar power plants over a year at half hourly period. The data are then aggregated, and scaled to obtain the set of weights for which the sum equals 1. Areas of interest are shown for one customer's load curve on Figure 1 [Left]. An example of the constructed

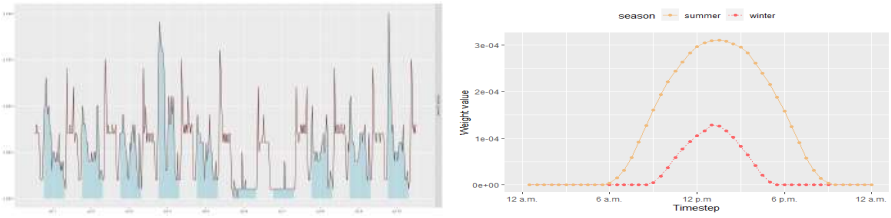


Figure 1: [Left] Load curve of one customer for ten days [Brown], areas highlighted [Blue] show the hours of sunlight, typically between 6 a.m. and 20 p.m. [Right] Solar weights for January 1 [Red] versus July 1 [Yellow]

weights is given on Figure 1 [Right]. They are higher in the summer than in the winter, as sun exposure is greater, and are equal to zero at night. We refer to them as solar weights in the rest of the paper, and denote them $(w_i^{sol})_{1\leq i\leq n}$.

Considering Y and $\hat{Y}$ are respectively a load curve and its prediction, both of size n. The solar Mean Absolute Error (MAE) $\mathcal{E}_{sol}$, is defined by:

$$\mathcal{E}_{sol}(Y,\hat{Y}) = \sum_{i=1}^{n} |Y_i - \hat{Y}_i| \times w_i^{sol}, \tag{1}$$

The solar MAE, introduced in (1), is used subsequently as the loss function to optimize during training of an autoencoder we apply to reduce the dimension of the load curves.

It is also used as our evaluation method to compare the models tested on the $C4$ testing subset.

The methodology is described in Section 2. Section 3 contains the application of the methods to EDF data.

2 Methodology for calculating prediction intervals

To reduce the dimension of the load curves, we apply an autoencoder as in Royer et al. (2020). The autoencoder takes the load curves X as inputs, and aims at reconstructing them first by encoding them into a lower dimensional space denoted I also called the latent space. Then it decodes the latent space I, using a decoding function d, back into the original space of the curves.

The autoencoder is first trained on the $C2$ subset of load curves, using the solar MAE, and fine-tuned on the $C4$ training subset. We denote $\mathcal{X} = (\mathbf{X}_k)_{1\leq k\leq m}$ the original load curves, $\mathcal{I} = (\mathbf{I}_k)_{1\leq k\leq m}$ the latent space, and m the number of observations.

Then, we model the latent space I using the customers' billing variables V. Considering the following multi-target nonlinear regression model:

$$\mathbb{E}(I|V) = f(V), \tag{2}$$

we estimate the regression function f from the $(\mathbf{I}_k, \mathbf{V}_k)_{1\leq k\leq m}$ observations using Bayesian neural networks. f is constituted of several layers. For instance, a one hidden layer feedforward neural network, with the variables V as the inputs and the latent space I as the outputs is:

$$\begin{aligned} \mathbf{h}_1 &= \sigma_1(\mathbf{W}_1^T \cdot V), \\ I &= \sigma_2(\mathbf{W}_2^T \cdot \mathbf{h}_1), \end{aligned} \tag{3}$$

h_1 is the output of the hidden layer, σ_p the nonlinear transformations, also called activation functions (here, ReLU function), $\mathbf{W}_p$ are the parameters of the neural network, and $1 \leq p \leq 2$. For a more in-depth review of neural networks and deep learning, refer to Goodfellow et al. (2016).

In a Bayesian setting, the parameters $\mathbf{W}_p$ follow a prior distribution, however in the context of neural networks, it is complicated to incorporate prior knowledge to this prior distribution. Hence, the chosen prior for the neural networks weights in our study is the standard multivariate normal distribution $\mathbf{W} \sim \mathcal{N}(0, \mathrm{Id})$, where Id is the identity matrix. The posterior distribution is :

$$p(\mathbf{W}|I, V) = \frac{\exp(-\frac{1}{2}\mathbf{W}^T\mathbf{W})p(I,V|\mathbf{W})}{p(I,V)}. \tag{4}$$

Variational inference is used to approximate the posterior distribution (see Blei et al. (2016)).

We consider the two following models:

- **BayesNN1**: a one hidden layer Bayesian neural networks, as described in (3).
- **BayesRN1**: a nine hidden layers Bayesian neural network, containing residual connections, as introduced by He et al. (2015), designed to prevent the vanishing gradient issue occurring when training deep neural networks.

They are trained on the $C2$ training subset and fine-tuned on the $C4$ training subset. From the estimated models, there are two ways for predicting the curve of a new customer using its estimated latent space. The first is described in Royer et al. (2020), where we have a strong industrial constraint to find the prediction amongst the library of existing curves. We do so by searching for the nearest neighbor's curve in the latent space $\mathcal{I}$. The second way relies on decoding the estimated latent space.

Assuming that, for a new customer, their billing variables $\mathbf{V}_{new}$ are known, the predictive distribution of $\hat{\mathbf{I}}_{new}$ is defined by:

$$p(\mathbf{I}_{new}|\mathcal{I},\mathcal{V},\mathbf{V}_{new}) \;\; = \int_{\mathcal{W}} p(\mathbf{I}_{new}|W,\mathbf{V}_{new})p(W|\mathcal{I},\mathcal{V})\,dW. \tag{5}$$

We simulate from $p(\mathbf{I}_{new}|\mathcal{I},\mathcal{V},\mathbf{V}_{new})$ a sample $\hat{\mathbf{I}}^{pos}_{new} = (\hat{\mathbf{I}}^{pos}_{new_j})_{j=1,\dots,J}$. Both methods described for predicting the load curve $\mathbf{X}_{new}$ are adapted to build the prediction intervals.

1. For each $j = 1,\dots,J$, we calculate the indexes of the nearest neighbors of $\hat{\mathbf{I}}^{pos}_{new_j}$:

$$\hat{k}_{I_j} = \underset{1\leq k\leq m}{\operatorname{argmin}} |\mathbf{I}_k - \hat{\mathbf{I}}^{pos}_{new_j}|. \tag{6}$$

   We denote $\mathcal{K} \subset \{1,\dots,m\}$ the set of values taken by the sample $(\hat{k}_{I_j})_{j=1,\dots,J}$ (i.e. the sample with no repetitions). We approximate the posterior distribution with the empirical distribution p_k on the indexes $k \in \mathcal{K}$. For each time $t = 1,\dots,n$, we compute the quantiles of the discrete distribution with support $\{X_k(t), k \in \mathcal{K}\}$ and of probability $(p_k)_{k\in\mathcal{K}}$.

2. We decode each $(\hat{\mathbf{I}}^{pos}_{new_j})_{j\in\{1,\dots,J\}}$ with the autoencoder:

$$\hat{\mathbf{X}}^{pos}_{new_j} = d(\hat{\mathbf{I}}^{pos}_{new_j}),\;\; j = 1,\dots,J. \tag{7}$$

   For each time $t = 1,\dots,n$, we compute the empirical quantiles of the sample $(\hat{\mathbf{X}}^{pos}_{new_j}(t))_{j\in\{1,\dots,J\}}$.

For both methods, we set a level of confidence of 80% and compute the 10th and 90th quantiles. Figure 2 provides a summary of the full methodology.

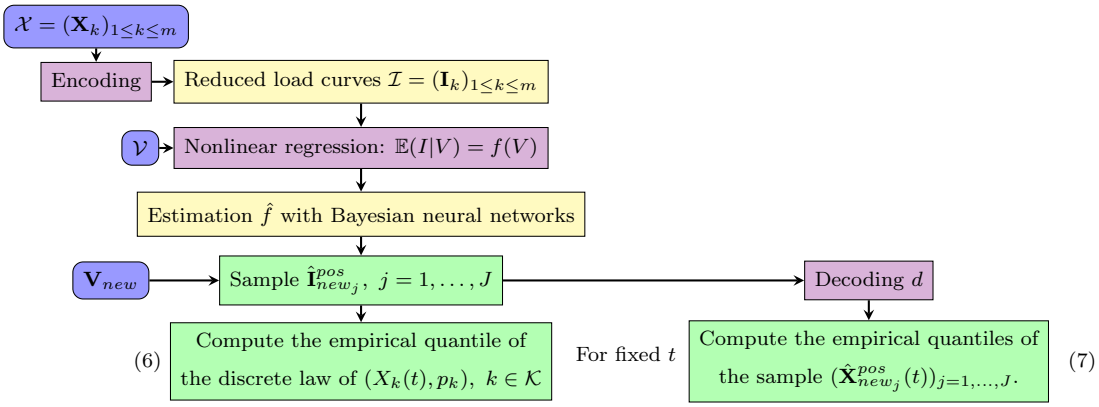


Figure 2: Outline of the two strategies for obtaining prediction intervals.

3 Results and discussion

Implementation is done using TensorFlow (Abadi et al. (2015)) and TensorFlow Probability (Dillon et al. (2017)).

Table 1 displays the median and mean errors obtained on the $C4$ testing subset when adopting either of the two methods: predicting the load curves, by searching for the nearest neighbor's load curve or decoding the predicted reduced load curve using the autoencoder, with $\mathbf{BayesNN}_1$ and $\mathbf{BayesRN}_1$. When decoding the samples, fine tuning improves the performances of both models. However, when searching for the nearest neighbor's load curve, fine tuning deteriorates the results. It is possible that decoding the samples with the autoencoder trained with fine tuning compensates these deterioration. The lowest errors are obtained in both cases with the $\mathbf{BayesRN}_1$ model, thus we show some results on the intervals obtained using this model.

Table 1: Prediction errors are displayed for both models on the $C4$ testing subset, for the two methods. The minimum is highlighted with an asterisk.

	$\mathcal{E}_{sol}(\mathbf{X}_{new}, \mathbf{X}_{\hat{k}_I})$				$\mathcal{E}_{sol}(\mathbf{X}_{new}, \hat{\mathbf{X}}_{new})$			
	Without fine-tuning		With fine-tuning		Without fine-tuning		With fine-tuning	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
$\mathbf{BayesNN}_1$	0.431	0.462	0.466*	0.499*	0.519*	0.583*	0.493	0.552
$\mathbf{BayesRN}_1$	0.409*	0.451*	0.503	0.559	0.572	0.639	0.464*	0.526*

Results obtained for credible intervals are shown on Figure 3. We see that proportion of half-hours for a given curve, for which the true value is within the intervals, obtained with the first method is quite high both with and without fine tuning, as in median it is over 0.85 (see Figure 3 [Left, Pink]). With the second method, this proportion is smaller, as the median lies between between 0.55 and 0.58 with and without fine-tuning.

Those results have to be put in perspective with the average length of intervals shown on Figure 3 [Right], as the average length of intervals obtained from the first method are wider than those obtained with the second method. Consequently, those intervals may be imprecise.

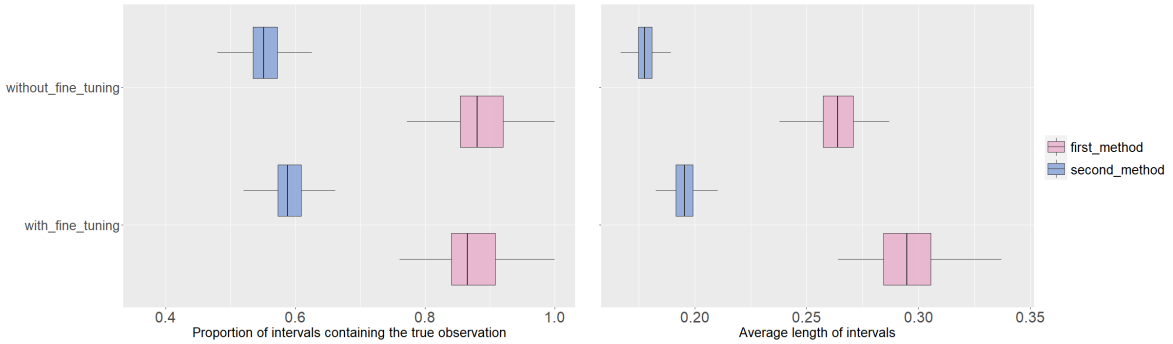


Figure 3: [Left] Proportions of the 80% prediction intervals containing the true observation on the $C4$ test subset, obtained with both methods. [Right] Average width of the 80% prediction intervals. From the samples of the posterior predictive distribution of the **BayesRN1** model (with fine-tuning [Bottom] and without fine-tuning [Top]), [Pink] Intervals obtained from the discrete distribution on $\hat{k}_I$ and [Blue] from decoding the samples.

Adding variables to the model relative to sizing of photo-voltaic installations could be a way to improve performances of the models. The decoding method for obtaining intervals is promising for future research, and may be applied for load profiling.

Bibliography

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112.518 : 859-877.

Dillon, J. V. et al. (2017). Tensorflow distributions. *arXiv preprint arXiv:1711.10604*.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*, MIT Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016), Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778.

Royer, H., Philippe, A., Charpentier, P. and Bozzi, L. (2020). Finding "twin" electrical load curves for new customers using deep learning, *Actes des Journées de Statistique de la Société Française de Statistique* 454–459.

Detecting the periodicity via an optimal testing procedure in integer-valued AR(p) process

Mohamed Djemaà SADOUN[1,2] & Mohamed BENTARZI[2]

mo-hamedsadoun@outlook.fr *mohamedbentarzi@yahoo.fr*

[1] *Laboratoire RECITS, BP 32 Bab Ezzouar, 16111-Algiers, Algeria*

[2] *Operational Research Department, University of USTHB, Algiers, Algeria*

Résumé. Ce travail est consacré principalement à l'étude de problème du test d'hypothèse concernant l'existence de la périodicité dans un processus autoregressive à valeurs entières d'ordre p, $(INAR(p))$, basé sur un opérateur d'amincissement binomial et conduit par une suite de variables aléatoires indépendantes suivant une distribution non spécifiée. Nous commençons d'abord par établir la propriété de normalité asymptotique locale (LAN), ensuite nous montrons la propriété de linéarité asymptotique locale concernant la suite centrale du processus sous-jacent. En utilisant ces résultats, nous construisons un test localement asymptotiquement optimal, pour l'hypothèse nulle d'un processus $(INAR(p))$ classique contre l'hypothèse alternative d'un processus périodiquement corrélé $(PINAR(p))$. Les performances du test établi sont montrées via une étude de simulation intensive et une application sur un ensemble de données réelles.

Mots-clés. Processus périodiquement corrélé $INAR(p)$, propriété (LAN), propriété de linéarité asymptotique, test optimal.

Abstract. This work is devoted mainly to the study of the periodicity testing problem in an integer-valued autoregressive $(INAR(p))$ process, based on binomial thinning operator, and driven by a sequence of independent random variables with an unspecified distribution. Local asymptotic normality (LAN) property is established, under some mild conditions. Moreover, the local asymptotic linearity property of its central sequence is verified. Using these results, we construct a locally asymptotically optimal (efficient) test, for the null hypothesis of classical $INAR(p)$ process against an alternative of periodically correlated $(PINAR(p))$ process. The performances of the established test are shown via an intensive simulation study and an application on real data sets.

Keywords. Periodically correlated $INAR(p)$ process, local asymptotic normality, asymptotic linearity property, optimal test.

1 Introduction and notations

The periodic integer-valued autoregressive $(PINAR)$ model has been introduced to model counting phenomena that evolve over time with a seasonal structure. The distribution of a parametric $PINAR(p)$ process is mainly described by two blocs of parameters, namely a periodic vector auto-regression coefficient and a periodic probability distribution on positive integers belonging to a parametric family, called an innovation distribution.

This work focuses on obtaining local asymptotic normality (LAN) property in order to test the presence of the periodicity in a p-order integer-valued autoregressive ($INAR\,(p)$) model.The periodic p-order integer-valued autoregressive ($PINAR_S\,(p)$) model is defined by the following difference stochastic equation:

$$y_t = \sum_{i=1}^{p} \varphi_{t,i} \circ y_{t-i} + \varepsilon_t,\ t \in \mathbb{Z}, \tag{1.1a}$$

where the underlying non-negative integer-valued process $\{y_t, t \in \mathbb{Z}\}$, is periodically correlated, in the sense of Gladyshev (1963) with period S (where $S \geq 2$, is a strictly positive integer, and the smallest positive integer such that $\varphi_{t+rS,i} = \varphi_{t,i}$, $r \in \mathbb{Z}_+$).
The innovation process, $\{\varepsilon_t, t \in \mathbb{Z}\}$, is a periodic sequence of independent non-negative integer-valued random variables, which is supposed to be independent of y_{t-i} and $\varphi_{t,i} \circ y_{t-i}$. Letting $\underline{\varphi}_s = \left(\varphi_{s,1}, \varphi_{s,2}, ..., \varphi_{s,p}\right)'$ be the the $p-$column vector of the auto-regression parameters, and $\underline{\alpha}_s = (\alpha_{s,1},\ \alpha_{s,2},\ ...,\ \alpha_{s,q})'$ be the $q-$column vector of the parameters of the innovation law. So, we define the $p+q-$column vector $\underline{\theta}_s = \left(\underline{\varphi}_s'; \underline{\alpha}_s'\right)' \in (0,1)^p \times A \subset \mathbb{R}_+^p \times \mathbb{R}_+^q$, $s = 1, 2, \ldots, S$, in order to define the global vector of the parameters of the model (1.1) of dimension $(p+q)\,S$, $\underline{\theta} = (\underline{\theta}_1'; \underline{\theta}_2';\ ...; \underline{\theta}_S')'$. Here $\{\varepsilon_t, t \in \mathbb{Z}\}$ is distributed with some discrete distribution belonging to the parametric family $\left\{\mathbb{G}_{\underline{\alpha}_s} \left| \underline{\alpha}_s = (\alpha_{s,1}, \alpha_{s,2}, ..., \alpha_{s,q})' \in A \subset \mathbb{R}_+^q \right.\right\}$, for $s = 1, 2, \ldots, S$, and where A is an open, convex subset of $\mathbb{R}_+^q$. Finally the symbol " $\circ$ " stands, as usual, the binomial thinning operator proposed by Steutel and Van Harn (1979), which is defined as follows:

$$\varphi_{t,i} \circ y_{t-i} = \begin{cases} \sum_{k=1}^{y_{t-i}} Y_{k,t,i}, & if\ y_{t-i} > 0, \\ 0, & if\ y_{t-i} = 0. \end{cases} \tag{1.1.b}$$

Note that the sequences of $(i.i.d)$ random variables of counts $\{Y_{k,t,i}, k \in \mathbb{N}, t \in \mathbb{Z}, i = 1, \ldots, p\}$ are mutually independents. $\{Y_{k,t,i}\}_{k \in \mathbb{N}, t \in \mathbb{Z}, i=1,...,p}$ are Bernoulli variables with periodic success probability $\varphi_{s,i} \in (0,1)$, $s = 1, 2, \ldots, S$ and $i = 1, \ldots, p$, which are independent of the innovation process.
Local structure of the parameters. Denoting by $H_g^{(n)}\,(\underline{\theta})$ a sequence of null hypotheses under which $\left\{y_t^{(n)}, t \in \mathbb{Z}\right\}$ is a sequence of realizations of an integer-valued process satisfying the model (1.1), with a $(p+q)-$vector parameters $\theta = \left(\underline{\varphi}; \underline{\alpha}'\right)' = \left(\varphi_1, \varphi_2, \ldots, \varphi_p; \alpha_1, \alpha_2,\ ..., \alpha_q\right)'$ and defining the $(p+q)\,S-$vector parameters $\underline{\theta} = \mathbf{1} \otimes \theta' = (\theta', \theta', \theta', ..., \theta')'$, where $\mathbf{1}$ is the $S-$dimensional vector $(1, 1, ..., 1)'$. Similarly, denoting $H_g^{(n)}\left(\underline{\theta}^{(n)}\right)$ the sequence of alternative hypotheses under which the sequence $\left\{y_t^{(n)}, t \in \mathbb{Z}\right\}$ is a sequence of realizations of a process satisfying the periodic integer-valued autoregressive model (1.1), with a $(p+q)\,S-$vector parameters $\underline{\theta}^{(n)} = \left(\theta_1^{(n)\prime};\ \theta_2^{(n)\prime};\ ...;\ \theta_S^{(n)\prime}\right)'$ where $\theta_s^{(n)} = \left(\underline{\varphi}_s^{(n)}; \underline{\alpha}_s^{(n)\prime}\right)'$ with $\underline{\alpha}_s^{(n)} = \left(\alpha_{s,1}^{(n)}, \alpha_{s,2}^{(n)},\ ..., \alpha_{s,q}^{(n)}\right)'$, $s = 1, ..., S$ where

$\varphi_{s,i}^{(n)} = \varphi_i + \dfrac{\delta_i^{(n)}}{\sqrt{n}} + \dfrac{\pi_{s,i}^{(n)}}{\sqrt{n}}$ and $\alpha_{s,j}^{(n)} = \alpha_j + \dfrac{\lambda_j^{(n)}}{\sqrt{n}} + \dfrac{h_{s,j}^{(n)}}{\sqrt{n}}$, $i = 1, 2, ..., p$, $j = 1, 2, ..., q$ and $s = 1, ..., S$, with the identification conditions $\pi_{S,i}^{(n)} = -\sum_{s=1}^{S-1} \pi_{s,i}^{(n)}$, $i = 1, ..., p$ and $h_{S,j}^{(n)} = -\sum_{s=1}^{S-1} h_{s,j}^{(n)}$, $j = 1, ..., q$, or equivalently $\underline{\theta}^{(n)} = \underline{\theta} + \mathbf{K}\, n^{-1/2} \underline{\boldsymbol{\tau}}^{(n)}$, where the $[(p+q)\, S] \times [(p+q)\, S]$ squared matrix $\mathbf{K}$ is given by

$$\mathbf{K} = \begin{pmatrix} \mathbf{I}_{(p+q)\times(p+q)} & & & \\ \vdots & & \mathbf{I}_{(p+q)(S-1)\times(p+q)(S-1)} & \\ \mathbf{I}_{(p+q)\times(p+q)} & & & \\ \mathbf{I}_{(p+q)\times(p+q)} & -\mathbf{I}_{(p+q)\times(p+q)} & \cdots & -\mathbf{I}_{(p+q)\times(p+q)} \end{pmatrix}_{[(p+q)S]\times[(p+q)S]},$$

where $\mathbf{I}_{(p+q)\times(p+q)}$ indicates the $(p+q)\times(p+q)$ identity matrix, and the $(p+q)\, S-$vector column $\underline{\boldsymbol{\tau}}^{(n)}$ is given by $\underline{\boldsymbol{\tau}}^{(n)} = \left(\underline{\boldsymbol{\Psi}}^{(n)\prime}; \underline{\mathbf{H}}_1^{(n)\prime}, ..., \underline{\mathbf{H}}_{S-1}^{(n)\prime}\right)'$ where

$$\underline{\boldsymbol{\Psi}}^{(n)\prime} = \left(\delta_1^{(n)}, \ldots, \delta_p^{(n)}, \lambda_1^{(n)}, ..., \lambda_q^{(n)}\right)' \in \mathbb{R}^{p+q}, \ \underline{\mathbf{H}}_s^{(n)} = \left(\pi_{s,1}^{(n)}, \ldots, \pi_{s,p}^{(n)}, h_{s,1}^{(n)}, ..., h_{s,q}^{(n)}\right)' \in \mathbb{R}^{p+q},$$

for $s = 1, ..., S-1$, such as $\sup_n \left(\underline{\boldsymbol{\tau}}^{(n)\prime}\, \underline{\boldsymbol{\tau}}^{(n)}\right) < \infty$. The global local perturbations of the parameters $\varphi_1, \varphi_2, ..., \varphi_p$; α_1, α_2, ..., α_q can be decomposed in to two component types, namely : the local non-periodic perturbations $\delta_1^{(n)}, \delta_2^{(n)}, ..., \delta_p^{(n)}$; $\lambda_1^{(n)}, \lambda_2^{(n)}, ..., \lambda_q^{(n)}$ and the quantities $\pi_{s,1}^{(n)}, \pi_{s,2}^{(n)}, ..., \pi_{s,p}^{(n)}$; $h_{s,1}^{(n)}, h_{s,2}^{(n)}, ..., h_{s,q}^{(n)}$, $s = 1, ..., S$, which can be interpreted as local periodic perturbations of the parameters $\varphi_1, \varphi_2, ..., \varphi_p$; $\alpha_1, \alpha_2, ..., \alpha_q$, respectively. Letting $\underline{\nu}^{(n)}$ be the $(p+q)\, S \times (p+q)\, S$ matrix given by $\underline{\nu}^{(n)} = \dfrac{1}{\sqrt{n}}\, \mathbf{K}$, one can easily rewrite the sequence of alternative hypotheses in the form $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)} \underline{\boldsymbol{\tau}}^{(n)}\right)$.

The present paper contributes to constructing while following the Le Cam (1960) methodology, an efficient parametric test of the existence of periodicity in the $INAR\,(p)$ process. This Le Cam's methodology allows for more precise and more general results under milder technical assumptions than, for instance, traditional Lagrangian multiplier testing procedures. The test we are obtaining is asymptotically valid under a large class of distributions, and locally asymptotically most stringent at some selected distribution.

2 Local asymptotic normality

Several researchers were interested in derivation of the LAN property for various time series models (see, Koul and Schick (1997), Linton (1993), Allal and Melhaoui (2006), Benghabrit and Hallin (1998), Bentarzi and Hallin (1996), and many others). To obtain the (LAN) property concerning our $PINAR_S\,(p)$ process, while following the same framework in Sadoun and Bentarzi (2021). We need the following definitions and notations.

Let $\underline{y}^{(n)} = \left(y_1^{(n)}, \ldots, y_n^{(n)}\right)$ be a realization of a finite size n of a stationary integer-valued process $\{y_t, t \in \mathbb{Z}\}$ satisfying (1.1a). Denote by $\Lambda_g^{(n)}\left(\underline{\theta}^{(n)}\right) = \Lambda_g^{(n)}(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)})$, the logarithm of the likelihood ratio of the calculated conditional likelihood $L_n\left(\underline{\theta}|y_0, \underline{y}^{(n)}\right)$ under $H_g^{(n)}(\underline{\theta})$ versus $L_n\left(\underline{\theta}^{(n)}|y_0, \underline{y}^{(n)}\right)$ under $H_g^{(n)}\left(\underline{\theta}^{(n)}\right)$, which is given by:

$$\begin{aligned}\Lambda_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right) &= \sum_{s=1}^{S}\sum_{\tau=0}^{m-1} \log P^{\theta_s^{(n)}}_{\left(y_{s-1+\tau S},\ldots,y_{s-p+\tau S}\right),y_{s+\tau S}} - \sum_{s=1}^{S}\sum_{\tau=0}^{m-1} \log P^{\theta_s}_{\left(y_{s-1+\tau S},\ldots,y_{s-p+\tau S}\right),y_{s+\tau S}} + o_P(1),\\ &= \sum_{s=1}^{S}\sum_{\tau=0}^{m-1} \log\left(\left(\left(\circledast_{i=1,\ldots,p} Bin_{y^*_{s-i+\tau S},\varphi^{(n)}_{s,i}}\right) \circledast \mathbb{G}_{\underline{\alpha}_s^{(n)}}\right)(y_{s+\tau S})\right)\\ &- \sum_{s=1}^{S}\sum_{\tau=0}^{m-1} \log\left(\left(\left(\circledast_{i=1,\ldots,p} Bin_{y^*_{s-i+\tau S},\varphi_{s,i}}\right) \circledast \mathbb{G}_{\underline{\alpha}_s}\right)(y_{s+\tau S})\right) + o_P(1),\end{aligned}$$

where $Bin_{y^*_{s-i+\tau S},\varphi_{s,i}}$, which stands as usual for the binomial distribution with parameters $\varphi_{s,i} \in (0,1)$ and $y^*_{s-i+\tau S} \in \mathbb{Z}_+$ with the point mass function $b_{y^*_{s-i+\tau S},\varphi_{s,i}}$, the notation $\left(\circledast_{i=1,\ldots,p} Bin_{y^*_{s-i+\tau S},\varphi_{s,i}}\right) \circledast \mathbb{G}_{\underline{\alpha}_s}$ indicates the convolution of the sum of $Bin_{y^*_{s-i+\tau S},\varphi_s}$ and $\mathbb{G}_{\underline{\alpha}_s}$.

Proposition 2.1. *Under some regularity conditions and under $H_g^{(n)}(\underline{\theta})$, we have:*
i) Taylor expansion in probability of $\Lambda_g^{(n)}\left(\underline{\theta}^{(n)}\right)$

$$\Lambda_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right) = \underline{\tau}^{(n)\prime}\underline{\Delta}^{(n)}(\underline{\theta}) - \frac{1}{2}\underline{\tau}^{(n)\prime}\Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})\,\underline{\tau}^{(n)} + o_P(1),$$

where the $(p+q)S \times (p+q)S$ square matrix $\Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})$ is the variance matrix of the score vector (also called central sequence) $\underline{\Delta}^{(n)}(\underline{\theta})$.
ii) Local Asymptotic Normality of the central sequence $\underline{\Delta}^{(n)}(\underline{\theta})$

$$\underline{\Delta}^{(n)}(\underline{\theta}) \xrightarrow{d} N_{(p+q)S}\left(\underline{0}, \Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})\right).$$

where $\Lambda_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right) = \log\left(dP^{(n)}_{\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}} \Big/ dP^{(n)}_{\underline{\theta}}\right)$ which represent the Radon-Nikodym derivative.is nonsingular and where $\Lambda_g^{(n)}(\underline{\theta} + \frac{1}{\sqrt{n}}\underline{\tau}^{(n)}) = \log\left(dP^{(n)}_{\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}} \Big/ dP^{(n)}_{\underline{\theta}}\right)$ which represent the Radon-Nikodym derivative.

3 Local asymptotic efficient test

Among the different consequence of the (LAN) property, we have the following result:

$$\begin{aligned}&\underline{\Delta}^{(n)}(\underline{\theta}) \Rightarrow N_{(p+q)S}\left(\underline{0}, \Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})\right) \text{ under } H_g^{(n)}(\underline{\theta}),\\ &\underline{\Delta}^{(n)}(\underline{\theta}) \Rightarrow N_{(p+q)S}\left(\Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})\,\underline{\tau}^{(n)}, \Gamma^{\underline{\Delta}^{(n)}}(\underline{\theta})\right) \text{ under } H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\tau}^{(n)}\right).\end{aligned}$$

Let us note $\eta = \Gamma^{\Delta}(\underline{\theta})\begin{pmatrix}\underline{\boldsymbol{\Psi}}\\ \underline{\mathbf{H}}\end{pmatrix}$, with $\underline{\boldsymbol{\Psi}}^{(n)} \to \underline{\boldsymbol{\Psi}}$ and $\underline{\mathbf{H}}^{(n)} \to \underline{\mathbf{H}}$ as $n \to \infty$, then the testing problem of the null hypothesis $H_g^{(n)}(\underline{\theta})$ versus the local alternative $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)$, i.e.,

testing a time-invariant $INAR(p)$ process, against a $S-$periodic $PINAR_S(p)$ process given by $(1.1a)$, becomes, quite simply, a testing problem tied to the experiment of Gaussian position. More precisely : testing the null hypothesis

$$H_{0,g} : N\left(\eta_0, \Gamma^{\Delta}(\underline{\theta})\right), \quad \left(\eta_0 = \Gamma^{\Delta}(\underline{\theta}) \begin{pmatrix} \underline{\Psi} \\ \underline{\mathbf{0}} \end{pmatrix}\right),$$

versus the alternative one

$$H_{1,g} : N\left(\eta, \Gamma^{\Delta}(\underline{\theta})\right), \quad \left(\eta = \Gamma^{\Delta}(\underline{\theta}) \begin{pmatrix} \underline{\Psi} \\ \underline{\mathbf{H}} \end{pmatrix}, \underline{\mathbf{H}} \neq 0\right).$$

The following proposition establish a locally asymptotically optimal test (so called most stringent test) to test $H_{0,g}$ versus $H_{1,g}$.

Proposition 3.1. *Under some regularity conditions, the test rejects the null hypothesis* $H_g^{(n)}(\underline{\theta})$ *whenever:*

$$\widehat{Q}_g^{(n)}\left(\widehat{\underline{\theta}}^{(n)}\right) = \widehat{\underline{\Delta}}^{(n)\prime}\left(\widehat{\underline{\theta}}^{(n)}\right)\left(\Gamma^{\underline{\Delta}^{(n)}}\left(\widehat{\underline{\theta}}^{(n)}\right)\right)^{-1}\widehat{\underline{\Delta}}^{(n)}\left(\widehat{\underline{\theta}}^{(n)}\right) > \chi^2_{(p+q)S,1-\alpha},$$

is such that :

(i) *has asymptotic level* α *under* $H_g^{(n)}(\underline{\theta})$,

(ii) *has asymptotic power:*

$$1 - \mathcal{F}\left(\chi^2_{1-\alpha}; (p+q)S; \underline{H}'\Gamma^{\underline{\Delta}^{(n)}}\left(\widehat{\underline{\theta}}^{(n)}\right)\underline{H}\right), \text{ under } H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right),$$

where $\mathcal{F}\left(\chi^2_{1-\alpha}; r; v\right)$ *denotes the non central chi-square distribution function with* r *degrees of freedom and non-centrality parameter* v;

(iii) *is locally asymptotically most stringent test against* $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)$.

4 Numerical illustration

The performances of the constructed efficient test are shown at first through a simulation study and reported in Table 1, where we have examined the power (and level) of our test by considering several time series with two different periods, $S = 2$ and $S = 4$. The periodic and the classical $INAR(1)$ data-generating processes are given for $s = 1, \ldots, S$ under the follows form : $\underline{\theta}_s = (\varphi_s, \underline{\alpha}_s)'$, $s = 1, .., S$

$$\begin{cases} y_{S\tau+s} = \varphi_s \circ y_{S\tau+(s-1)} + \varepsilon_{S\tau+s} & \text{with } \varepsilon_{S\tau+s} \rightsquigarrow \mathcal{P}(\underline{\alpha}_s), \\ y_{S\tau+s} = \varphi_s \circ y_{S\tau+(s-1)} + \varepsilon_{S\tau+s} & \text{with } \varepsilon_{S\tau+s} \rightsquigarrow \mathcal{G}(\exp(-\underline{\alpha}_s)). \end{cases}$$

Indeed, a periodic $INAR_2(1)$ process, $M1$, and a periodic $INAR_4(1)$ process, $M2$, are used to simulate time series of small, moderate and relatively large sizes ($n = 100, 150, 200, 300,$ and $400, 600, 800$). The sets of parameter values are chosen such that the underlying models are periodically stationary, where the periodically stationary condition applies here is the sufficient condition : $\varphi_s \leq 1, s = 1, \ldots, S$ and $\prod_{s=1}^{S}\varphi_s < 1$, and for each model cited above we use two types of distributions for the innovation process $\{\varepsilon_t\}_{t\in\mathbb{Z}_+}$, namely *Poisson* $\mathcal{P}(\alpha_t)$ and *Geometric* $\mathcal{G}(e^{-\alpha_t})$, where for Geometric case α_t is chosen such as $\alpha_t \simeq e^{-x}, x \in \mathbb{N}^*$. The true parameter values of these models are:

Model $M1$ $\underline{\theta} = [\underline{\theta}_1'; \underline{\theta}_2']' = [(.9, 4); (.2, 1)]'$,
$\underline{\theta} = [\underline{\theta}_1'; \underline{\theta}_2']' = [(.9, \exp(-4)); (.2, \exp(-1))]'$.
Model $M2$ $\underline{\theta} = [\underline{\theta}_1'; ...; \underline{\theta}_4']' = [(.9, 6); (.1, 1); (.6, 3); (.4, 4.5)]'$,
$\underline{\theta} = [\underline{\theta}_1'; ...; \underline{\theta}_4']' = [(.9, \exp(-6)); (.1, \exp(-1)); (.6, \exp(-3)); (.4, \exp(-4.5))]'$.

Table 1
Empirical powers and levels of the efficient tests ϕ for the level 5%

	n	100	150	200	300	400	600	800
		ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ
$M1$	$g_{(\alpha_t,\mathcal{P})}$	.8917	.9694	.9920	.9994	1	1	1
	$g_{(\alpha_t,\mathcal{G})}$	.9594	.9937	.9995	1	1	1	1
$M2$	$g_{(\alpha_t,\mathcal{P})}$	.9214	.9597	.9832	.9975	.9995	1	1
	$g_{(\alpha_t,\mathcal{G})}$	.9469	.9771	.9904	1	1	1	1

For the real data example, we consider the seasonal data set of 365 observations, consisting on the daily counts of daytime road accidents in Schiphol area, in Netherlands for the year (2001). The logical variables d_α which equal to 1 each time we reject H_0, the powers and the sizes of the test $\phi_\alpha^1, \phi_\alpha^2$ for each level α are given in Table 2 below as follows:

Table 2
Statistics values, powers and sizes for $PINAR_S(1)$ models

	α	$\chi_2^*(\mathcal{P})$	$\chi_2^2(\mathcal{P})$	$d_\alpha(\mathcal{P})$	$\phi_\alpha^1(\mathcal{P})$	$\phi_\alpha^2(\mathcal{P})$	$\chi_2^*(\mathcal{G})$	$\chi_2^2(\mathcal{G})$	$d_\alpha(\mathcal{G})$	$\phi_\alpha^1(\mathcal{G})$	$\phi_\alpha^2(\mathcal{G})$
$S=5$	$\alpha = 5\%$	13.85	15.51	0	.3410	.0547	13.16	15.51	0	.4498	.0562
	$\alpha = 10\%$	13.63	13.36	1	.4913	.0111	13.55	13.36	1	.6033	.0974
$S=7$	$\alpha = 5\%$	33.67	21.03	1	.9281	.0547	27.03	21.03	1	.9316	.0562
	$\alpha = 10\%$	39.65	18.54	1	.9620	.0111	29.89	18.55	1	.9667	.0974

Bibliographie

Allal, J. and Melhaoui, S. (2006). Optimal detection of exponential component in autoregressive models. *Journal of Time Series Analysis.* 27 (6) : 793 − 810.

Bentarzi, M. and Hallin, M. (1996). Locally Optimal Tests Against Periodic Autoregression. *Econometric Theory* 12 pp. 88 − 112.

Benghabrit, Y. and Hallin, M. (1998). Locally asymptotically optimal Tests for $AR(p)$ against diagonal bilinear dependence. *Journal of Statistical Planning and Inference* 68, pp. 47 − 63.

Koul, H.L. and Schick, A. (1997). Efficient estimation in non-linear autoregressive time series models. *Bernoulli.* 3, pp. 247 − 277.

Le Cam. L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics.* 3 : 37 − 98.

Linton, O. (1993). Adaptive Estimations in $ARCH$ Models. *Econometric Theory.* 9(4) pp. 539 − 569.

Sadoun, M. and Bentarzi, M. (2021). Locally asymptotically efficient estimation for parametric $PINAR(p)$ models. *Statistica Neerlandica.* DOI: 10.1111/stan.12234, 1 − 33.

Détection de ruptures dans les signaux EMG de l'activité musculaire du trapèze supérieur.

Nassim Sahki [1] & Anne Gégout-Petit [1] & Sophie Wantz-Mézières [1]

[1] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France ;*
nassim.sahki@inria.fr ; anne.gegout-petit@univ-lorraine.fr ;
sophie.mezieres@univ-lorraine.fr

Résumé. Nous nous intéressons à l'étude du développement de la myalgie du muscle trapèze en milieu de travail. Les données recueillies sont des signaux EMG (électromyographique) de l'activité musculaire du trapèze supérieur de 30 participants en bonne santé, sans cervicalgie chronique, effectuant différentes activités informatiques lors d'une journée expérimentale. Dans le but de la détection de rupture, nous nous plaçons dans le cadre de la détection séquentielle où l'on admet que le signal EMG arrive point par point, en temps réel. Ensuite, nous considérons la statistique CUSUM basée sur le score pour détecter des changements de régime dans l'activité musculaire. Notre méthodologie de détection prend en charge l'estimation des paramètres (moyenne et variance du régime pré-changement) d'une manière online sur le signal EMG. Les résultats de détection ont permis ensuite de caractériser les différents types d'activités.

Mots-clés. Activité du muscle trapèze, signal temporel EMG, détection séquentielle de rupture, statistique CUSUM basée sur le score.

Abstract. We are interested in studying the development of myalgia of the trapezius muscle in the workplace. The data collected are EMG (electromyographic) signals of upper trapezius muscle activity from 30 healthy participants without chronic neck pain, performing different computer activities on an experimental day. For the purpose of breaking detection, we place ourselves within the framework of the sequential detection where we admit that the EMG signal arrives point by point, in real time. Next, we consider the score-based CUSUM statistic to detect diet changes in muscle activity. Our detection methodology supports the estimation of the parameters (mean and variance of the pre-change regime) in an online way on the EMG signal. The detection results then made it possible to characterize the different types of activities.

Keywords. Trapezius muscle activity, EMG time signal, Sequential change-point detection, CUSUM statistic based-score.

1 Contexte

Dans ce travail, nous proposons d'appliquer une méthodologie de détection de ruptures on-line sur des données réelles de santé. Les données de santé ont été fournies par

l'institut de recherche INRS de Nancy. Elles concernent les signaux temporels EMG (électromyographiques) de l'activité musculaire du trapèze supérieur de 30 sujets effectuant différentes activités bureautiques au cours d'une journée expérimentale. Le recueil des données s'inscrit dans le cadre d'une étude sur le risque de développement de la myalgie du muscle trapèze en milieu de travail (Veiersted et al [1] ; Goudy et al. [2]).

La méthodologie de détection séquentielle de ruptures requiert le choix d'une statistique récursive, un seuil de détection et une règle d'arrêt. Dans notre application, nous utilisons la version semi-paramétrique de la statistique classique des sommes cumulées (CUSUM) basée sur la fonction de score (Page [3] ; Tartakovsky et al. [4]). Pour le choix du seuil de détection, nous utilisons la méthode analytique basée sur les inégalités de Wald (1945) pour fixer un seuil constant (Sahki et al. [5]). Et quant à la règle d'arrêt, nous utilisons la règle corrigée (Sahki et al. [6]) de la règle adoptée basiquement par la procédure CUSUM, dans le but de minimiser les fausses alarmes. La règle d'arrêt corrigée consiste à signaler une détection si seulement si la statistique de CUSUM dépasse le seuil pendant un temps c prédéterminé.

La procédure de détection séquentielle basée sur le score, exige la connaissance des paramètres de moyenne μ_0 et de variance σ_0^2 du régime pré-changement, et la définition de l'objectif de détection (type et niveau de la rupture que l'on souhaite détecter). Nous proposons dans ce travail, un algorithme de détection séquentielle de ruptures permettant d'estimer les paramètres du régime pré-changement (μ_0, σ_0^2) d'une façon online sur le signal EMG. A chaque instant, l'algorithme permet de détecter une rupture d'augmentation ou de diminution pour le type (moyenne et/ou variance) et le niveau de la rupture recherchés.

Notre objectif est la détection de changements de régime dans l'activité musculaire du trapèze, et la caractérisation des différentes activités bureautiques effectuées durant la journée expérimentale.

2 Données électromyographiques - EMG

L'activité musculaire du trapèze est évaluée par la technique de l'électromyographie de surface, permettant de quantifier l'activité électrique des muscles qu'entraînent les contractions musculaires volontaires.

Dans le but de faciliter le processus de détection, nous avons effectué un traitement du signal, qui comprend une transformation logarithmique et un lissage en calculant la moyenne par lots successifs de $R = 4$ points (lissage sur 1 seconde). A la fin du lissage, une activité de 50 minutes couvre 3000 points. Nous présentons sur la Figure 1 le signal EMG lissé de la première activité du sujet 1.

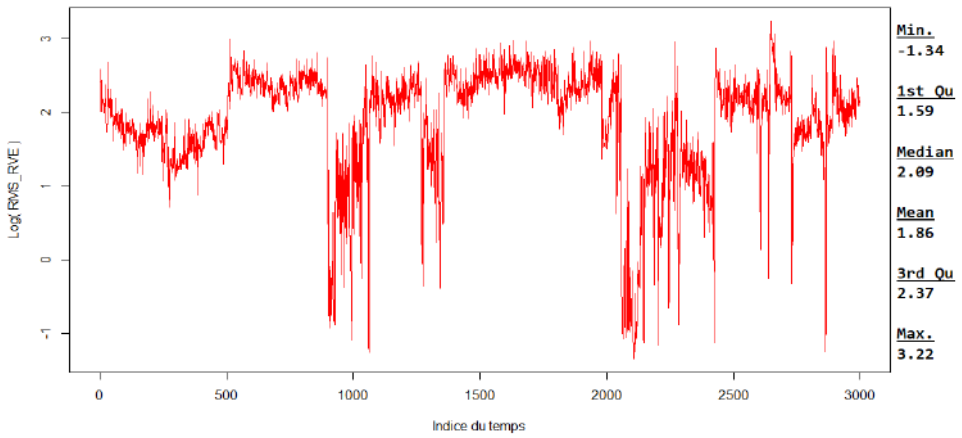


FIGURE 1 – Signal EMG lissé de la première activité du sujet 1.

3 Détection de changements d'activité musculaire

Soit $X_1, ..., X_t$ l'échantillon associé au signal EMG (RMS-RVE) séquentiellement observée jusqu'à l'instant t.

3.1 Statistique, seuil de détection et règle d'arrêt

Tartakovsky et al. [4] proposent une fonction de score $S(X_1, ..., X_t)$ calculée en fonction des observations, pour la détection de rupture sur la moyenne et/ou la variance. Elle est définie par :

$$S_t(\delta, q) = C_1 \cdot Y_t + C_2 \cdot Y_t^2 - C_3, \tag{1}$$

où $Y_t = (X_t - \mu_0)/\sigma_0$ sont les données centrées et standardisées à l'instant t sous le régime pré-changement, et

$$C_1 = \delta \cdot q^2, \qquad C_2 = \tfrac{1-q^2}{2}, \qquad C_3 = \tfrac{\delta^2 \cdot q^2}{2} - \log(q),$$
$$\text{avec} \quad \delta = (\mu_1 - \mu_0)/\sigma_0 \quad \text{and} \quad q = \sigma_0/\sigma_1.$$

Les paramètres δ et q sont respectivement la différence de moyenne et le ratio de variance que l'on voudrait détecter. Ils sont définis en fonction de notre objectif de détection.

La statistique CUSUM (Page [3]) basée sur le score est définie de manière récursive au temps t comme suit :

$$W_t(\delta, q) = \max\{0,\ W_{t-1}(\delta, q) + S_t(\delta, q)\}, \quad t \geq 1; \qquad W_0(\delta, q) = 0. \tag{2}$$

La W-statistique (2) est calculée en fonction de la moyenne et de la variance dans le régime pré-changement (μ_0, σ_0^2) et de l'objectif de détection $(\delta,\ q)$.

Notons que sans changement de régime, le score S_t a une espérance négative alors qu'elle devient positive après changement. C'est pourquoi la règle d'arrêt consiste à signaler une détection lorsque la statistique de test dépasse un seuil.

Dans l'application, nous utilisons l'inégalité de Wald [7] pour déterminer le seuil constant de Wald $h^W(\alpha)$. Il est donné après avoir fixé le taux instantané de fausse alarme toléré α, en respectant :

$$h^W(\alpha) = -\ln(\alpha). \tag{3}$$

La procédure corrigée signale l'existence d'une rupture lorsque la statistique de détection dépasse le seuil de détection pendant un temps $c \geq 1$, c étant un paramètre à fixer à l'avance ($c = 1 \Leftrightarrow$ règle d'arrêt classique). Le temps de la détection est donné par :

$$T^c_{h^W(\alpha)} = \min_{t \geq 1}\{t + c - 1;\ \bigcap_{i=t}^{t+c-1} (W_i(\delta, q) \geq h^W(\alpha))\} \tag{4}$$

3.2 Estimation des paramètres de chaque régime

L'algorithme doit inclure la période de l'estimation des paramètres pré-ruptures (μ_0, σ_0^2) au début du traitement et après chaque détection. Voici les détails sur cette méthode d'estimation de μ_0 et σ_0^2 :

- La moyenne μ_0 est estimée au début et réestimée après chaque détection sur la partie du signal de longueur L points.
- La variance σ_0^2 est estimée par une méthode pseudo-bayesienne comme suit :
  Soit $\widehat{s}_0^2$ l'estimateur de la variance sur les L premiers points du signal. Et $\widehat{s}_k^2$ l'estimateur de la variance sur les L points du signal suivant la $k^{\text{ème}}$ détection. A la $k^{\text{ème}}$ détection, l'estimateur de σ_k^2 est calculée itérativement comme suit :

$$\widehat{\sigma}^2_{0,0} = \widehat{s}_0^2$$

$$\widehat{\sigma}^2_{0,k} = \frac{\widehat{\sigma}^2_{0,k-1} + \widehat{s}_k^2}{2}, \qquad k = 1, 2, ...$$

Dans la pratique, nous devons fixer à priori un ensemble de paramètres pour pouvoir appliquer l'algorithme de détection. Cet ensemble comprend le choix de l'objectif de détection (q, δ), du temps d'attente de la règle d'arrêt corrigée (c) et la longueur des données d'estimation (L).

A chaque transition, l'algorithme de détection comprend d'abord une période d'estimation sur un temps L, puis la procédure de détection de rupture. Lorsque la statistique dépasse le seuil et que la rupture est signalée, au temps T, nous considérons à posteriori que le régime a changé en $T - c$, de sorte nous estimons les paramètres (μ_0 et σ_0^2) de ce nouveau régime à partir de $T - c$. Nous montrons un exemple dans Figure 2. Dans celui-ci, $L = 90$, $c = 100$. Et on signale l'alarme au temps $T = 346$, ce qui correspond à un changement de régime au temps $Cpt = 246$.

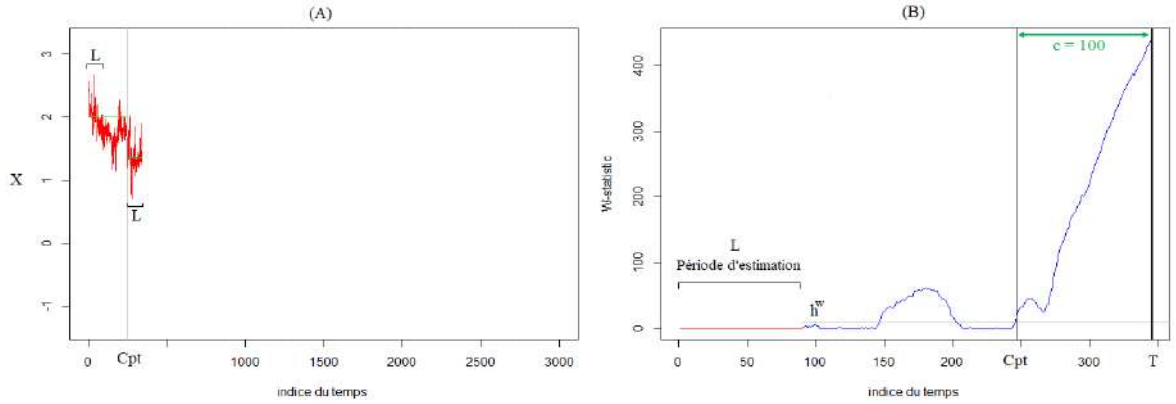


FIGURE 2 – Exemple de détection sur le signal EMG d'une activité. L'objectif de détection est sur la moyenne ($q = 1$) pour un niveau $\sigma_0\delta = 0.69$. (A) : le signal dans le premier régime et la période d'estimation $L = 90$ dans le deuxième régime, après la détection de diminution à l'instant $Cpt = 246$. Les moyennes estimées dans le premier et second régime sont données par les segments verts. (B) : W-statistique calculée dans le premier régime (ligne bleue), et signalement de la rupture produite à l'instant $Cpt = 246$ au temps d'arrêt $T = 346$.

3.3 Résultats

Nous présentons dans la Figure 3, un exemple de détection sur une activité entière du sujet 1. Sur cet exemple, nous avons détecté 14 ruptures de moyenne (8 d'augmentation et 6 de diminution).

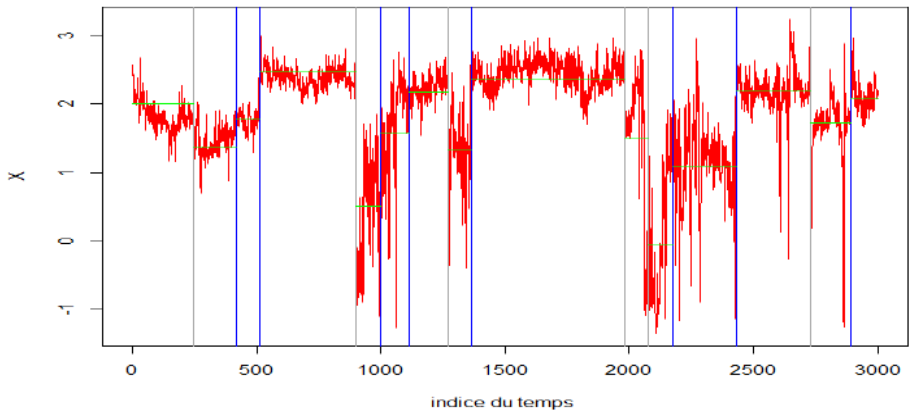


FIGURE 3 – Résultats de détection sur la première activité du sujet 1. Les ruptures d'augmentation (resp. de diminution) sont marquées par des lignes verticales bleues (resp. grises). Les lignes horizontales vertes sont les moyennes estimées dans chaque régime.

Les résultats de détection de la Figure 3 ont été obtenus en utilisant le Seuil Constant de Wald (Seuil-CW). Nous présenterons lors de la communication les résultats de détection en utilisant le seuil instantané empirique dynamique (Seuil-IED) que nous avons proposé dans Sahki et al. [5]. Dans ce papier, nous avons présenté des des résultats de détection sur des données simulées.

Les résultats de détection dans chaque activité de chaque sujet ont permis la caractérisation des différents types d'activités en fonction du nombre et de l'amplitude des ruptures. Nous présenterons les résultats lors de la communication.

Références

[1] Kaj Bo Veiersted, Mikael Forsman, Gert-Åke Hansson, and Svend Erik Mathiassen. Assessment of time patterns of activity and rest in full-shift recordings of trapezius muscle activity–effects of the data processing procedure. *Journal of Electromyography and Kinesiology*, 23(3) :540–547, 2013.

[2] N Goudy and L McLean. Using myoelectric signal parameters to distinguish between computer workers with and without trapezius myalgia. *European journal of applied physiology*, 97(2) :196–209, 2006.

[3] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2) :100–115, 1954.

[4] Alexander G Tartakovsky, Aleksey S Polunchenko, and Grigory Sokolov. Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1) :4–11, 2012.

[5] Nassim Sahki, Anne Gégout-Petit, and Sophie Wantz-Mézières. Performance study of change-point detection thresholds for cumulative sum statistic in a sequential context. *Quality and Reliability Engineering International*, 36(8) :2699–2719, 2020.

[6] Nassim Sahki, Anne Gégout-Petit, and Sophie Wantz-Mézières. Détection statistique de rupture dans le cadre online. In *JdS 2019-51èmes Journées de Statistique*, 2019.

[7] Abraham Wald. Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, 16(2) :117–186, 1945.

[8] Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blažek, and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9) :3372–3382, 2006.

Space-time trends detection and dependence modelling approaches by functional Peaks-Over-Thresholds. Application to precipitation in Burkina Faso

Béwentaoré Sawadogo[1,2] & Liliane Bel[1] & Diakarya Barro[2,3]

[1] *Université Paris Saclay, INRAe, AgroParisTech, UMR MIA-Paris, 75005, Paris, France, liliane.bel@agroparistech.fr*
[2] *LANIBIO, UJKZ, UFR-SEA, BP: 7021, Ouagadougou 03, Burkina Faso, sbewentaore@yahoo.fr*
[3] *UFR-SEG, Université Thomas SANKARA, 12 BP: 417 Ouagadougou 12, Burkina Faso, dbarro2@gmail.com*

Abstract. In this paper, we propose a method for estimating trends in extreme spatio-temporal processes using both information from marginal distributions and dependence structure. We combine two statistical approaches of extreme value theory allowing on the one hand to handle temporal and spatial non-stationarities via a tail trend function with a spatio-temporal structure in the marginal distributions and by modelling on the other hand the dependence structure by a latent stationary process using generalized ℓ-Pareto processes. This methodology for trend analysis of extreme events is applied to precipitation data from Burkina Faso. We show that a significant increasing trend for the 50 and 100 years return levels in some parts of the country.
Keywords: Non-stationary POT, Generalized ℓ-Pareto process, Space-time Extremes, Dependence Modelling, Trends detection, Climate Change.

Résumé. Dans cet article, nous proposons une méthode pour estimer les tendances dans les extrêmes de processus spatio-temporels en utilisant à la fois des informations provenant des distributions marginales et de la structure de dépendance. Nous combinons deux approches statistiques de la théorie des valeurs extrêmes permettant d'une part de prendre en compte les non-stationnarités temporelles et spatiales via une fonction de tendance de queue à structure spatio-temporelle dans les distributions marginales et en modélisant d'autre part la dépendance spatiale à l'aide d'un processus stationnaire latent en utilisant les processus ℓ-Pareto généralisés. Cette méthodologie d'analyse des tendances des événements extrêmes est appliquée aux données de précipitations du Burkina Faso. Nous montrons que les niveaux de retour à 50 et 100 ans montrent une croissance significative dans certaines zones du pays.
Mots-clés: POT non stationnaire, processus ℓ-Pareto généralisé, extrêmes spatio-temporels, dépendance spatiale, détection de tendance, changement climatique.

1 Introduction

Climate extremes occur in a non stationary framework due to the action of climate change. To study extremes in this framework, the Extreme value Theory initially designed for independent or stationary processes must be extended. Several works focus on that topic, some of them present transform a non-stationary time series into a stationary series for which classical EVT can be applied ([1],[7]). A frequently alternative for addressing non-stationarity in EVT models is to include covariates, mostly space and time, to the model the parameters ([2],[11]). These approaches work marginally and do not take spatial dependence into account. Recently some forms of non-stationary dependence structures, have been studied by [8],[10], in the framework of spatial max-stable processes. Although attractive, these spatial models are very computationally expensive for large dimensions. And alternative for max-stable modelling is the modelling of threshold exceedances. The spatial Peak Overs Threshold (POT) approach introduced in a stationary framework by [6], and generalized by [5],[4] that have defined the family of generalized ℓ-Pareto process are good candidates for modelling the spatial dependence structure of extremes of spatio-temporal processes.

In this paper we propose a methodology coupling two statistical approaches of EVT, non-stationarity is handled in the marginal distributions via a trend function with spatio-temporal structure([7]) while the spatial dependence structure is modelled using a functional POT model([3, 4]) to obtain a flexible spatio-temporal method of threshold exceedances to assess presence of trends in the extremes. We aim at evaluating trend in extreme precipitation in sub-Saharan Africa by these means, and we illustrate theses evolution by representing non stationary high return levels over the area.

2 Methodology

2.1 Space-time trends modelling

Let $X = \{X_t(s), s \in S, t \in T\}$ be a continuous non-stationary space-time stochastic process with sample paths in the family of continuous functions $\mathcal{C}(S \times T)$ equipped with the uniform norm $\| \, . \, \|_\infty$, where $S \times T \subset \mathbb{R}^d \times \mathbb{R}^+$ and $\mathcal{C}_+(S \times T)$ its restriction to non-negative functions deprived of the null function. In practice X is observed at each stations $s_1, \cdots, s_m$ and at dates $t = 1, \cdots, n$. Let $F_{t,s}$ be the continuous univariate marginal distribution with a common right endpoint x_F and $Z = \{Z(s), s \in S\}$ an unobserved latent stationary stochastic process with sample paths in $\mathcal{C}(S \times T)$ satisfying the proportional tail condition such that

$$\lim_{x \to x_F} \frac{P\left(X_t(s) > x\right)}{P\left(Z(s) > x\right)} = c_\theta \left(\frac{t}{n}, s\right), \text{with}, \frac{1}{m} \sum_{j=1}^{m} \int_0^1 c_\theta \left(u, s_j\right) du = 1, \quad u \in [0,1], \tag{1}$$

where $c_\theta : [0,1] \times S \longrightarrow]0, \infty[$ is assumed to be a continuous and positive function depending on a parameter vector $\theta \in \Theta \subset \mathbb{R}$, called tail trend function or skedasis function ([6, 7]). The skedasis function describes the evolution of extreme events jointly in space and time. Moreover we assume that the continuous marginal distributions F_Z of the latent process has the same right endpoint x_F. F_Z is in the maximum domain of attraction condition for some constant $\gamma \in \mathbb{R}$ and appropriate real normalization constants $a_Z > 0$, $u_Z \in \mathbb{R}$. Thanks to equation(1) and the convergence of $\{Z(s), s \in S\}$ exceedances to a GPD distribution we deduce a pseudo-sample of $\{Z_t(s)\}$ from observations of $\{X_t(s), s \in S, t \in T\}$ in the following manner ([7]):

$$\widehat{Z}_t(s_j) = \left\{ \widehat{c_\theta}\left(\frac{t}{n}, s_j\right) \right\}^{-\widehat{\gamma}} \left[X_t(s_j) - \frac{\left\{ \widehat{c_\theta}\left(\frac{t}{n}, s_j\right) \right\}^{\widehat{\gamma}} - 1}{\widehat{\gamma}} \left(\widehat{a}_Z - \widehat{\gamma}\widehat{u}_Z\right) \right], \quad (2)$$
$$j = 1, \cdots, m \; ; \; t = 1, \cdots, n,$$

where $\widehat{\gamma}, \widehat{a}_Z, \widehat{u}_Z$ and $\widehat{c_\theta}$ are respectively consistent estimators of γ, a_Z, u_Z and c_θ which we will discuss later in the section (2.2).

The modelling is then focused on the evaluation of the extreme spatial dependence structure in Z using functional POT([5],[3, 4]). In the multivariate and spatial framework a threshold exceedance for a random function $Z = \{Z(s), s \in S\}$ is defined by [5] to be an event of the form $\{\ell(Z) > u\}$ for some $u \geq 0$, where $\ell : \mathcal{C}(S) \longrightarrow \mathbb{R}^+$ is a continuous and homogeneous non-negative risk function. The risk function ℓ can be for instance the maximum, minimum, average or value at a specific point $s_0 \in S$. As in ([6]), we assume that the stationary process Z is a general functional regular variation process. Under minimal assumptions on the risk function and for n large enough, the conditional distribution of ℓ-exceedance for some threshold $u \geq 0$ of the normalized process can be approximated by a generalized ℓ-Pareto process, i.e for some a_n and b_n functions ([4])

$$P\left\{ \left[\frac{Z - b_n}{\ell(a_n)} \right] \in A \mid \ell\left(\frac{Z - b_n}{\ell(a_n)}\right) \geq u \right\} \longrightarrow P\{W_\ell \in A\}, n \longrightarrow \infty, \quad (3)$$

where W_ℓ is a non-degenerate stochastic process over S and belongs to the family of generalized ℓ-Pareto processes with tail index γ.

2.2 Statistical inference

Marginal parameters of the process γ and a_n of the latent stationary process Z can be estimated by maximizing the independence log-likelihood taking $\widehat{u}_Z = \ell(b_n) = q_{1-\alpha}\{\ell(Z)\}$ a high quantile of $\ell(Z)$. Assuming the dates of exceeding the marginal thresholds $\widehat{u}_Z$ form a Poisson process, its density is the trend function parameter c_θ which is estimated using maximum likelihood method, assuming a parametric log-linear model as in ([9]).

The inference on the spatial dependence structure of the latent stationary process, is driven by modelling the angular component of ℓ-Pareto process ([5], [3]) by a log-Gaussian

process with stationary increments within the framework of a Brown-Resnick model. The parameters of the isotropic semi-variogram are estimated using the gradient scoring rule method ([3]).

The non-stationary m-period return levels of the process X at each location $x_m(s)$ are evaluated inverting the equation(2) from return levels calculated from Z, $z_m(s)$.

$$x_m(s) = z_m(s)\widehat{c}_\theta(t_m, s)^{\widehat{\gamma}} + \frac{\widehat{c}_\theta(t_m, s)^{\widehat{\gamma}} - 1}{\widehat{\gamma}} \left(\widehat{a}_Z - \widehat{\gamma}\widehat{u}_Z\right), \quad s \in S, \tag{4}$$

$t_m = 1 + \dfrac{n_x m}{n}$ with n_x the number of days in the year and n the size of the sample observed. $z_m(s)$ are calculated at each site by extrapolating the marginal parameters on the whole area.

3 Main Results

This study uses time series of daily precipitation measurements from 1957 to 2016 provided by ten synoptic stations extracted from the Burkina Faso climatological database. These stations have been selected to ensure good spatial uniformity and representativeness of different climatic regimes and data quality. In order to limit the problems related to seasonal rainfall cycles, on each station we worked from the sub-series corresponding to rainy days that is the period from May to October. Figure (1) shows the estimated trend function c_θ by a log-linear model for the ten stations. The frequencies of extreme precipitation are increasing in areas such as Ouahigouya, Bogande, Boromo, Gaoua, and Po. On the other hand, at the Ouagadougou and Fada stations, extreme rainfall frequencies tend to decrease ($\theta < 0$). Figure (2) shows the non-stationary return levels $x_m(s)$

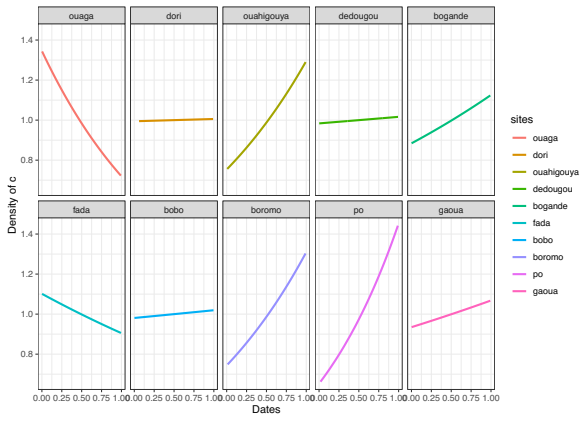


Figure 1: Local adjustment of evolution of the frequencies of extreme precipitation by a log-linear model trend c_θ.

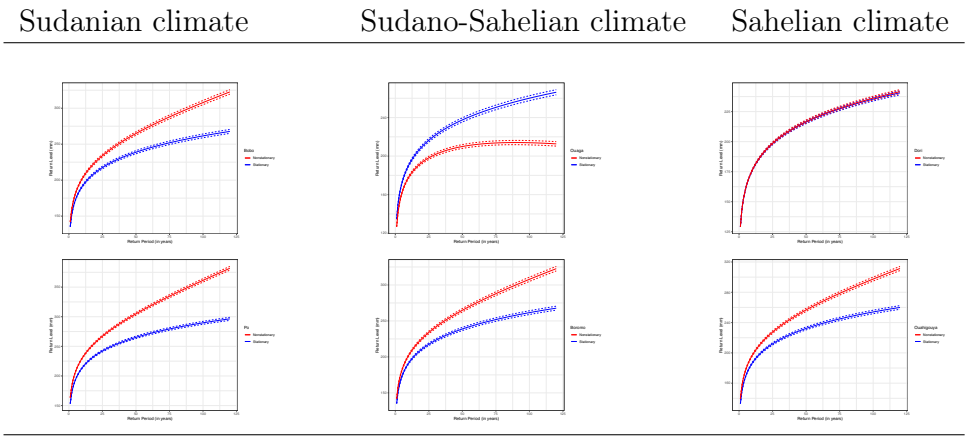


Figure 2: Trends of stationary and non-stationary return levels at reference stations calculated with a log-linear tail trend function.

together with the stationary return levels $z_m(s)$ for six reference stations.

The map of non-stationary return level for a return period $m = 50$ and 100 years are displayed in Figures (3). The extreme precipitation are likely to be observed on average at least once every 50 years (resp. 100 years), will be particularly intense in the Sudanian and Sudano-Sahelian zone and less intense in the Sahelian zone, with a potentially quite strong spatial dependence within a radius of 200 km. The southwest and eastern region of the country will be most affected by extreme precipitation.

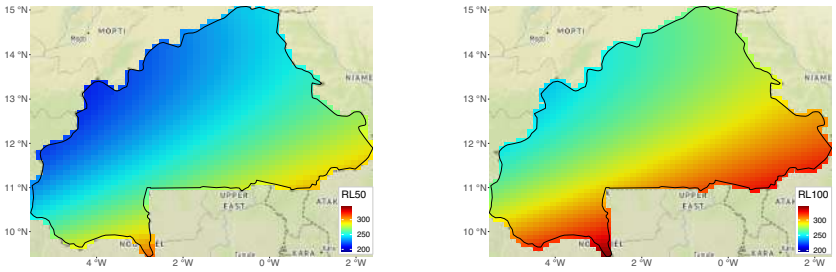


Figure 3: Maps of the non-stationary 50-years return levels(left) and 100-years return level (right) obtained by extrapolating the information from the log-linear tail trend function and the dependence structure.

4 Conclusion

In this study we proposed a new flexible methodology for trend detection in the extremes capable of capturing marginal non-stationarities and the dependence structure between margins using generalized ℓ-Pareto processes. We calculated the non-stationary return

levels of precipitation in Burkina Faso and exhibit some regions in which there may be a significant increasing or decreasing of extreme precipitation.

References

[1] Didier Dacunha-Castelle, Thi Thu Huong Hoang, and Sylvie Parey. Modeling of air temperatures: preprocessing and trends, reduced stationary process, extremes, simulation. *Journal de la Société Française de Statistique*, 156(1):138–168, 2015.

[2] Anthony C Davison and Richard L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990.

[3] R. de Fondeville and A. C. Davison. High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592, 2018.

[4] R. de Fondeville and A.C. Davison. Functional peaks-over-threshold analysis, 2020. Preprint.

[5] Clément Dombry and Mathieu Ribatet. Functional regular variations, pareto processes and peaks over threshold. *Statistics and its Interface*, 8(1):9–17, 2015.

[6] Ana Ferreira and Laurens De Haan. The generalized pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737, 2014.

[7] Ana Ferreira, Petra Friederichs, Laurens de Haan, Cláudia Neves, and Martin Schlather. Estimating space-time trend and dependence of heavy rainfall. *arXiv preprint arXiv:1707.04434*, 2017.

[8] Raphaël Huser and Marc G Genton. Non-stationary dependence structures for spatial extremes. *Journal of agricultural, biological, and environmental statistics*, 21(3):470–491, 2016.

[9] Aline Mefleh, Romain Biard, Clément Dombry, and Zaher Khraibani. Trend detection for heteroscedastic extremes. *Extremes*, 23(1):85–115, 2020.

[10] Gilles Nicolet, Nicolas Eckert, Samuel Morin, and Juliette Blanchet. Assessing climate change impact on the spatial dependence of extreme snow depth maxima in the french alps. *Water Resources Research*, 54(10):7820–7840, 2018.

[11] Sylvie Parey, Thi Thu Huong Hoang, and Didier Dacunha-Castelle. Different ways to compute temperature return levels in the climate change context. *Environmetrics*, 21(7-8):698–718, 2010.

Extension du modèle logistique conditionnel pour la modélisation de la dynamique d'action de la pepsine lors de la digestion

Ousmane Suwareh [1], David Causeur [2] & Françoise Nau [3]

[1] *1STLO, INRAE, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France, ousmane.suwareh@agrocampus-ouest.fr*
[2] *IRMAR UMR6625, CNRS, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France, david.causeur@agrocampus-ouest.fr*
[3] *1STLO, INRAE, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France, francoise.nau@agrocampus-ouest.fr*

Résumé. La digestion des protéines est un processus dynamique complexe au cours duquel les molécules de protéines sont progressivement fragmentées en peptides sous l'action d'enzymes telles que la pepsine. Cette enzyme pourrait couper préférentiellement certaines séquences peptidiques plutôt que d'autres selon les propriétés physico-chimiques de ces séquences. Les dispositifs expérimentaux *in vitro* permettent de reproduire la digestion, afin d'identifier par la suite les peptides libérés à différents temps d'observation à l'aide de la spectrométrie de masse. Cependant, cette approche ne permet pas de déduire de manière certaine qu'un peptide a été coupé ou non par la pepsine entre deux temps d'observation. L'estimation d'un modèle de régression de la probabilité de coupure d'une séquence par la pepsine doit tenir compte de cet étiquetage incertain de la variable réponse. Nous proposons un algorithme d'estimation de type Expectation-Maximization, consistant à alterner l'estimation du modèle de régression et la correction adaptative des valeurs de la variable réponse. En application à des données expérimentales, nous démontrons que cet algorithme conduit à de meilleures estimations des modèles de regression, et une meilleure interprétabilité biologique des résultats issus de la sélection de variables explicatives au sein du profil de propriétés physico-chimiques.

Mots-clés. Digestion, Données incomplètes, Peptidomique, Regression logistique conditionnelle, Spéctrométrie de masse.

Abstract. Proteins digestion is a complex and dynamic process during which protein molecules are progressively reduced into peptides through the action of enzymes such as pepsin. This enzyme may preferentially cleave some peptide sequences rather than others depending on the physico-chemical properties of these sequences. Experimental *in vitro* systems make it possible to reproduce digestion in order to subsequently identify the peptides released at different observation times using mass spectrometry. However, this approach does not allow to state with certainty on the cleavage or not of a peptide by pepsin between two observation times. The estimation of a regression model of the probability for a sequence to be cleaved by pepsin must consider this uncertain labelling of

the response variable. We propose an Expectation-Maximization algorithm which consists in alternating the estimation of the regression model and the adaptive correction of the response variable values. In application to experimental data, we show that this algorithm leads to better estimates of regression models, and a better biological interpretability of the results resulting from the selection of explanatory variables within the profile of physico-chemical properties.

Keywords. Digestion, Mislabeled data, Peptidomics, Conditional logistic regression, Mass spectrometry.

1 Introduction

Les dispositifs expérimentaux de digestion *in vitro* permettent de simuler les processus de digestion en environnement contrôlé, les aliments cibles et les enzymes d'intérêt étant choisis, divers paramètres des phases gastriques et intestinales étant fixés. Dans le cas d'aliments protéiques, la spectrométrie de masse permet d'observer l'action des enzymes en identifiant les fragments de protéine, les peptides, libérés au cours du temps. Ces peptides sont des séquences de résidus d'acides aminés résultant de la coupure par une enzyme d'une séquence plus grande, définissant ainsi une filiation entre peptides. L'objectif de ces dispositifs expérimentaux est de comprendre la dynamique de génération par l'enzyme de sous-séquences en cascade, à partir de coupures successives des protéines introduites en début d'expérience. En particulier, une des questions majeures est celle de la spécificité ou non de l'action de la pepsine (protéase de la phase gastrique). On entend ici par spécificité le fait que la coupure d'une séquence de résidus d'acides aminés n'interviendrait pas au hasard mais répondrait à des lois physico-chimiques favorisant une coupure à certains endroits plutôt que d'autres.

D'un point de vue statistique, on peut voir cette question comme un problème d'estimation d'un modèle de régression liant la probabilité de coupure d'une séquence de résidus d'acides aminés à des propriétés physico-chimiques de cette séquence décrites par un profil de variables explicatives. A titre d'illustration, ce profil de variables explicatives peut contenir le nombre de résidus d'acides aminés de la séquence, le nombre de résidus d'acides aminés chargés ou encore un indice d'hydrophobie.

2 Modèle de régression logistique conditionnel

Si s_{ijk} désigne une séquence de résidus d'acides aminés observée au temps t_j et pouvant être clivée au temps t_k, avec $t_j \in \{0.5, 2, 5, 10, 20, 30\}$, $t_k \in \{2, 5, 10, 20, 30, 60\}$ et $t_k > t_j$, alors $X(s_{ijk}) = (X_1(s_{ijk}), \ldots, X_p(s_{ijk}))'$ est le profil de variables explicatives de la séquence s_{ijk}, de dimension $p \geq 1$. Par ailleurs, $Y(s_{ijk})$ est la variable réponse qui prend la valeur 1, si la séquence s_{ijk} a été coupée par l'enzyme entre t_j et t_k, 0 sinon. Des études

préalables incitent à proposer le modèle de régression logistique additif suivant, reliant la probabilité $\pi(s_{ijk}) = \mathbb{P}(Y(s_{ijk}) = 1 \mid X(s_{ijk}) = x(s_{ijk}))$ au profil $x(s_{ijk})$ observé pour s_{ijk}, dans lequel les effets marginaux de chaque variable explicative sont décrits de manière non-paramétrique :

$$\text{logit}\big(\pi(s_{ijk})\big) \quad = \quad \beta_0^{(ijk)} + \sum_{r=1}^{p} f_r^{(ijk)}\big(x_r(s_{ijk})\big), \tag{1}$$

où $\beta_0^{(ijk)} \in \mathbb{R}$ et $f_r^{(ijk)}$ est une fonction décrivant l'effet marginal de $X_r(s_{ijk})$ sur $\pi(s_{ijk})$.

On fait l'hypothèse que les profils de variables explicatives restent les mêmes au cours du temps mais que les paramètres et fonctions d'effet du modèle (1) peuvent être différents selon l'intervalle $[t_j; t_k]$ sur lequel on modélise la probabilité de coupure. On estime donc de manière séparée et indépendante les modèles de coupure aux différents intervalles d'observation.

Pour une séquence s_{ijk} de résidus d'acides aminés observée au temps t_j, le fait qu'aucune sous-séquence de s_{ijk} ne soit observée à t_k sera dans la suite considérée comme une condition suffisante pour affirmer que s_{ijk} n'a pas été coupée par l'enzyme entre t_j et t_k et donc que $Y(s_{ijk}) = 0$. En revanche, si au moins une sous-séquence de s_{ijk} est observée à t_k, l'affirmation que $Y(s_{ijk}) = 1$ repose le plus souvent sur une incertitude. En effet, les sous-séquences de s_{ijk} observées à t_k peuvent aussi résulter de la coupure par l'enzyme d'autres séquences également observées à t_j. L'estimation du modèle (1) repose donc sur des observations partielles de la variable réponse $Y(s_{ijk})$, pour toutes les séquences s_{ijk} observées au temps t_j.

Cette situation particulière de données manquantes s'apparente à celle introduite par Breslow *et al.* (1978) pour une problématique d'estimation de modèle de survie en épidémiologie. Sous l'hypothèse qu'une seule séquence parmi les candidates pourrait avoir généré les sous-séquences observées, le modèle de régression logistique conditionnel proposé par Breslow *et al.* (1978) serait approprié. Par ailleurs, plus récemment, Hung *et al.* (2017) introduisent également un modèle de régression logistique dans lequel seule une version approximative de la variable réponse est observée. Dans le cas présent, on propose d'aborder l'estimation du modèle (1) comme un problème d'estimation en situation de données manquantes, le mécanisme générant l'absence de données étant non-ignorable.

3 Estimation itérative

En première approximation, on peut choisir de considérer que si au moins une sous-séquence de s_{ijk} est observée à t_k alors $Y(s_{ijk}) = 1$. Cette approximation conduit sans ambiguïté à des valeurs observées de $Y(s_{ijk})$ pour toutes les séquences s_{ijk} et permet d'en

déduire une estimation du modèle (1) par la méthode usuelle du maximum de vraisemblance.

Pour aller au-delà de cette première approximation, toujours dans le cas où au moins une sous-séquence de s_{ijk} est observée à t_k, on identifie la liste exhaustive $\mathcal{C}(s_{ijk}) = \{c_1(s_{ijk}), \ldots, c_K(s_{ijk})\}$ des séquences candidates à la coupure, toutes observées à t_j, et sur lesquelles l'action de l'enzyme a pu générer les sous-séquences de s_{ijk} observées à t_k. La séquence s_{ijk} est donc un élément particulier de $\mathcal{C}(s_{ijk})$ et, par convention, on considèrera que $c_1(s_{ijk}) = s_{ijk}$. Même si la notation ne le laisse pas paraître par souci de simplicité, $K = K(s_{ijk})$ n'est pas le même selon la séquence s_{ijk}.

Si ce nombre K de séquences candidates est plus grand que 1, on ne peut pas affirmer avec certitude que $Y(s_{ijk}) = 1$, mais seulement qu'il existe $c \in \mathcal{C}(s_{ijk})$ tel que $Y(c) = 1$. On en déduit la probabilité conditionnelle $\pi(s_{ijk} \mid \mathcal{C}(s_{ijk}))$ que $Y(s_{ijk}) = 1$:

$$\begin{aligned} \pi(s_{ijk} \mid \mathcal{C}(s_{ijk})) &= \mathbb{P}\big(Y(s_{ijk}) = 1 \mid \cup_{c \in \mathcal{C}(s_{ijk})} [Y(c) = 1]\big), \\ &= \frac{\mathbb{P}\big(Y(s_{ijk}) = 1\big)}{\mathbb{P}\big(\cup_{c \in \mathcal{C}(s_{ijk})} [Y(c) = 1]\big)}. \end{aligned}$$

Par souci de simplicité, la notation ci-dessus omet le conditionnement $\cap_{c \in \mathcal{C}(s_{ijk})}[X(c) = x(c)]$ par les profils de variables explicatives des séquences candidates.

Sous l'hypothèse d'indépendance entre les coupures par la pepsine des séquences candidates, on obtient :

$$\pi(s_{ijk} \mid \mathcal{C}(s_{ijk})) = \frac{\pi(s_{ijk})}{1 - \prod_{c \in \mathcal{C}(s_{ijk})} \big(1 - \pi(c)\big)}.$$

L'observation des séquences de résidus d'acides aminés au temps t_k et la connaissance préalable des paramètres du modèle (1) permettent donc d'estimer la probabilité que la séquence s_{ijk} ait été coupée entre t_j et t_k. On propose ici d'actualiser la valeur observée de $Y(s_{ijk})$ selon la règle suivante : si $\pi(s_{ijk} \mid \mathcal{C}(s_{ijk})) \geq 0.5$, alors $Y(s_{ijk}) = 1$ et $Y(s_{ijk}) = 0$ sinon. Après actualisation des valeurs obervées de la variable réponse pour toutes les séquences observées à t_j, l'estimation du modèle (1) peut elle-même être actualisée. L'alternance d'actualisation des valeurs observées de la variable réponse et de l'estimation du modèle se répète ainsi jusqu'à convergence.

4 Illustration

On se contente ici de présenter quelques résultats illustrant l'application de la méthode présentée ci-dessus à des données expérimentales visant à décrire l'action de la pepsine au cours de la digestion de protéines. La Table 1 donne les nombres de séquences considérées comme coupées ou non par la pepsine entre 30s et 2mn de digestion, pour chacun des

trois cycles nécessaires pour estimer le modèle de la probabilité de coupure à partir d'un profil de 38 variables explicatives.

	Peptides non coupés	Peptides probablement coupés	Peptides probablement non coupés	Peptides coupés
Initialement	584	147	0	14
Après 1er cycle	584	92	55	14
Après 2ème cycle	584	81	66	14
Après 3ème cycle	584	75	72	14

Table 1: Valeurs observées de la variable réponse après chaque itération de l'algorithme.

Pour 147 séquences peptidiques parmi les 745 identifiées à 30s, les données ne permettaient pas d'affirmer avec certitude qu'elles avaient été coupées par la pepsine avant 2mn de digestion. Au bout de 3 cycles, un peu moins de 49% de ces 147 séquences ont finalement été identifiées comme étant non coupées. Au delà du 3ème cycle, il n'y a plus d'évolution des valeurs observées de la variable réponse.

Tous les modèles de coupure ont ainsi été améliorés, au sens d'un critère d'Akaike du modèle final plus faible. Les modèles présentent également une complexité moins importante que lors de la première estimation, avec des courbes d'effets marginaux désormais plus interprétables biologiquement. A titre d'exemple, La Figure 1 montre l'estimation initiale et finale, après trois itérations, de la fonction décrivant l'effet marginal du point isoélectrique d'un peptide sur la probabilité de coupure par la pepsine en début de digestion (entre 30s et 2mn après le début). L'estimation finale révèle un effet plus marqué que celle obtenue par l'approximation initiale.

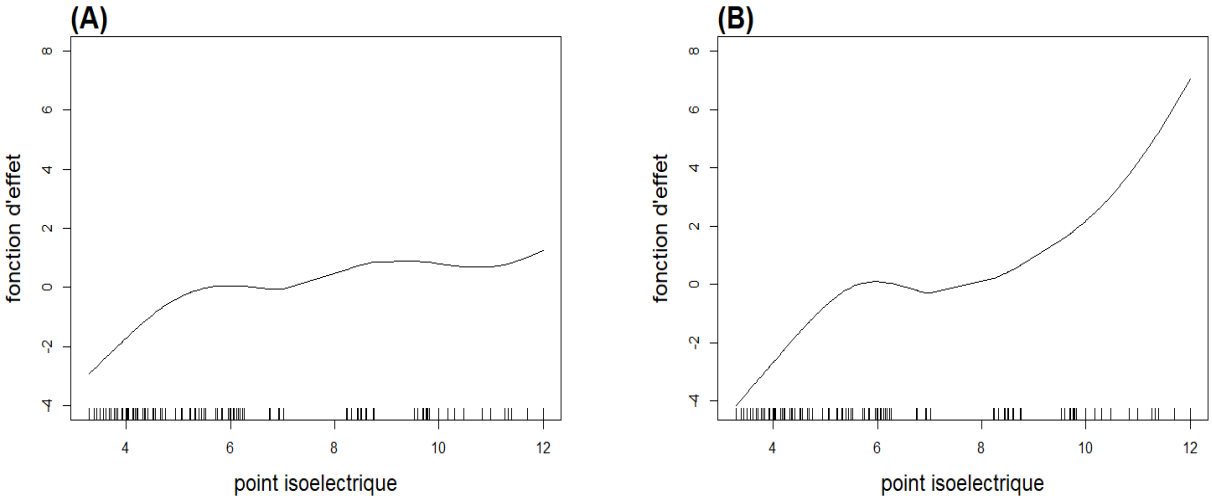


Figure 1: Courbe partielle du modèle pour le point isoelectrique (A) sans ajustement des valeurs observées de la variable réponse et (B) après 3 cycles d'actualisation/estimation

Globalement, une valeur élevée du point isoelectrique d'un peptide, donc chargé positivement, favoriserait sa coupure par la pepsine. Ce résultat est cohérent avec la présence de deux résidus d'acides aminés fortement chargés négativement dans le site actif de la pepsine (Sepulveda *et al.*, 1975), qui pourrait donc avoir une affinité particulière pour les peptides chargés positivement.

Bibliographie

Prentice, R. L. and Breslow, N. E. (1978) Retrospective studies and failure time models, *Biometrika*, Volume 65, Issue 1, Pages 153–158.
Hung, H., Jou, Z.-Y. and Huang, S.-Y. (2018), Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74: 145-154.
Sepulveda, P., Marciniszyn, J., Liu, D. and Tang, J., (1975). Primary structure of porcine pepsin. III. Amino acid sequence of a cyanogen bromide fragment, CB2A, and the complete structure of porcine pepsin. *J. Biol. Chem.* 250, 5082–5088.

Testing a class of time varying CHARN MODELS

Youssef SALMAN [1,2] & Joseph NGATCHOU WANDJI [1] & Zaher KHRAIBANI [2]

[1] *IECL, Lorraine University, France*
[2] *LM, Lebanese University, Faculty of sciences, Lebanon*
Email: youssef.salman@univ-lorraine.fr
Email: joseph.ngatchou-wandji@univ-lorraine.fr
Email: Zaher.khraibani@ul.edu.lb

Abstract. We study a likelihood ratio test for testing a general class of CHARN models. The LAN property is established for the family of likelihoods under study. The test is proved to be optimal.

Keywords. Times series, changepoint, likelihood ratio test, LAN, optimality.

Résumé. Nous étudions un test du rapport de vraissemblance pour tester une classe générale de modèles CHARN. La propriété LAN est établie pour la famille des vraissemblances considérées. L'optimalité du test est prouvée.

Mots-clés. Séries chronologiques, rupture, test du rapport de vraisemblance, LAN, optimalité.

1 Introduction

We consider $X_1, \ldots, X_n$, observations generated by the CHARN model "Conditional Heteroskedastic Autoregressive Nonlinear" [1]

$$X_t = T(\rho_0 + \gamma \odot \omega(t); \mathbf{X}_{t-1}) + V(\mathbf{X}_{t-1})\varepsilon_t, t \in \mathbb{Z} \tag{1}$$

where $(X_t)_{t\in\mathbb{Z}}$ is a locally stationary ergodic process, $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a standard white noise with knowing density function f, $\mathbf{X}_t = (X_t, \ldots, X_{t-d+1})^\top$, T and V are real smooth functions and $V > 0$, $\rho_0^\top \in \mathbb{R}^p$, $\gamma = \left(\gamma_1^\top, \ldots, \gamma_{k+1}^\top\right)^\top$ and for every $j = 1, \ldots, k+1$, $\gamma_j \in \mathbb{R}^p$, $\omega(t) = (\mathbb{1}_{[\tau_0,\tau_1)}(t), \mathbb{1}_{[\tau_1,\tau_2)}(t), ..., \mathbb{1}_{[\tau_{k-1},\tau_k)}(t), \mathbb{1}_{[\tau_k,\tau_{k+1})}(t))^\top = (\omega_1, \ldots, \omega_{k+1}) \in \{0,1\}^{k+1}$. We assume that, for every $j = 1, \ldots, k$, $n_j(n)$ represents the number of observations in $[\tau_{j-1}, \tau_j)$, $\tau_0 = 1 < \tau_1 < \cdots < \tau_{k+1} = n$. We suppose that, as $n \to +\infty$, $n_j(n) \to +\infty$ and $\frac{n_j(n)}{n} \to \alpha_j$. F_j is the disribution function of X_j on $[\tau_{j-1}, \tau_j)$. For every $\Pi = \left(\Pi_1^\top, \ldots, \Pi_{k+1}^\top\right)^\top \in \mathbb{R}^{p(k+1)}$ and $\Theta = \left(\Theta_1^\top, \ldots, \Theta_{k+1}^\top\right) \in \mathbb{R}^{p(k+1)}$,

$$\Pi \odot \Theta = \Pi_1\Theta_1 + \cdots + \Pi_{k+1}\Theta_{k+1}$$

and for any β and $\tilde{\beta} \in \mathbb{R}^{p(k+1)}$,

$$\beta \circ \widetilde{\beta} = \widetilde{\beta} \circ \beta = \begin{pmatrix} \begin{pmatrix} \beta_{1,1} \\ \vdots \\ \beta_{1,p} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \beta_{k+1,1} \\ \vdots \\ \beta_{k+1,p} \end{pmatrix} \end{pmatrix} \circ \begin{pmatrix} \begin{pmatrix} \widetilde{\beta}_{1,1} \\ \vdots \\ \widetilde{\beta}_{1,p} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \widetilde{\beta}_{k+1,1} \\ \vdots \\ \widetilde{\beta}_{k+1,p} \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \beta_{1,1}\widetilde{\beta}_{1,1} \\ \vdots \\ \beta_{1,p}\widetilde{\beta}_{1,p} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \beta_{k+1,1}\widetilde{\beta}_{k+1,1} \\ \vdots \\ \beta_{k+1,p}\widetilde{\beta}_{k+1,p} \end{pmatrix} \end{pmatrix} \in \mathbb{R}^{p(k+1)}$$

stands for the Hadamard product [2].
We aim to test

$$H_0 : \gamma = \gamma_0 \quad \text{against} \quad H_1^{(n)} : \gamma = \gamma_n = \gamma_0 + \beta/\sqrt{n}, \quad \text{where } \gamma_0 \text{ and } \beta \in \mathbb{R}_*^{p(k+1)} \tag{2}$$

In this purpose, we used the likelihood ratio test.
Our primary goal is to verify that $\tau_j,\ j = 1, \ldots, k$, are instants of change or not. This work is preliminary to the construction of a method for testing weak changes.
The changepoint theory was started by Page [3]. He used the cumulative sum (CUSUM) to detect the changepoints in the mean of independent observations. Since then, there has been a lot of work on changepoints see, eg. [4], [5], [6], [7].

2 Asymptotics

In the present paper, we investigate the case where the functions T, V and f are known, as well as the nuisance parameter ρ_0.

2.1 Likelihood ratio test

The log-likelihood ratio test Θ_n for H_0 against $H_1^{(n)}$, can be expressed as follow

$$\Theta_n = \Theta_{1n} - \Theta_{2n} + o_P(1), \tag{3}$$

where

- $\Theta_{1n} = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{t=1}^{n} \left\{ \frac{1}{V(\mathbf{X}_{t-1})} \beta^\top N(\gamma_0, \mathbf{X}_{t-1}) \phi_f[\varepsilon_t(\gamma_0)] \right\}$
- $\Theta_{2n} = \dfrac{1}{2n} \displaystyle\sum_{t=1}^{n} \left\{ \frac{1}{V^2(\mathbf{X}_{t-1})} \beta^\top M(\gamma_0, \mathbf{X}_{t-1}) \beta \phi_f'[\varepsilon_t(\gamma_0)] - \frac{1}{V(\mathbf{X}_{t-1})} \beta^\top \mathcal{H}(\gamma_0, \mathbf{X}_{t-1}) \beta \phi_f[\varepsilon_t(\gamma_0)] \right\}$

- $M(\gamma_0, \mathbf{X}_{t-1}) = N(\gamma_0, \mathbf{X}_{t-1}) N^\top(\gamma_0, \mathbf{X}_{t-1})$
- $N(\gamma_0, \mathbf{X}_{t-1}) = \omega(t) \circ D_\gamma\big[T(\gamma_0, \mathbf{X}_{t-1})\big]$
- $D_\gamma[T(\gamma_0, \mathbf{X}_{t-1})] = \Big(\nabla_{\gamma_1}[T(\gamma_0, \mathbf{X}_{t-1})], \ldots, \nabla_{\gamma_{k+1}}[T(\gamma_0, \mathbf{X}_{t-1})]\Big)^\top \in \mathbb{R}^{p(k+1)}$
- $\nabla_{\gamma_i}[T(\gamma_0, \mathbf{X}_{t-1})] = \left(\dfrac{\partial T}{\partial \gamma_{i,1}}(\gamma_0, \mathbf{X}_{t-1}), \dfrac{\partial T}{\partial \gamma_{i,2}}(\gamma_0, \mathbf{X}_{t-1}), \ldots, \dfrac{\partial T}{\partial \gamma_{i,p}}(\gamma_0, \mathbf{X}_{t-1})\right)^\top \in \mathbb{R}^p$ is the gradient of T with espect to γ_i
- $\mathcal{H}(\gamma_0, \mathbf{X}_{t-1}) = \omega(t)\omega(t)^\top \circ H_\gamma(\gamma_0, \mathbf{X}_{t-1})$
- $\mathcal{H}(\widetilde{\gamma}, \mathbf{X}_{t-1}) = \begin{pmatrix} \mathcal{H}_1(\widetilde{\gamma}, \mathbf{X}_{t-1}) & 0 & \ldots & 0 \\ 0 & \mathcal{H}_2(\widetilde{\gamma}, \mathbf{X}_{t-1}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \mathcal{H}_{k+1}(\widetilde{\gamma}, \mathbf{X}_{t-1}) \end{pmatrix} \in \mathbb{R}^{p(k+1)\times p(k+1)}$
- $\mathcal{H}_i(\widetilde{\gamma}, \mathbf{X}_{t-1}) = \omega_i^2 \begin{pmatrix} \dfrac{\partial^2 T}{\partial \gamma_{i,1}^2}(\widetilde{\gamma}, \mathbf{X}_{t-1}) & \ldots & \dfrac{\partial^2 T}{\partial \gamma_{i,p}\partial \gamma_{i,1}}(\widetilde{\gamma}, \mathbf{X}_{t-1}) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 T}{\partial \gamma_{i,1}\partial \gamma_{i,p}}(\widetilde{\gamma}, \mathbf{X}_{t-1}) & \ldots & \dfrac{\partial^2 T}{\partial \gamma_{i,p}^2}(\widetilde{\gamma}, \mathbf{X}_{t-1}) \end{pmatrix} \in \mathcal{M}_p(\mathbb{R})$ for $i = 1, \ldots, k+1$.

where

$$\varepsilon_t(\gamma) = \frac{X_t - T(\rho_0 + \gamma \odot \omega(t), \mathbf{X}_{t-1})}{V(\mathbf{X}_{t-1})}.$$

2.2 LAN property

After finding the expression of the likelihood ratio test, we must find its distribution. To compute the distribution of the test under the null hypothesis, we need to establish the LAN property stated in [5]. To do this, we proved that, under H_0,

$$\Theta_{2n} \underset{n\to+\infty}{\longrightarrow} \frac{\eta(\gamma_0, \beta)}{2} \tag{4}$$

and

$$\Theta_{1n} \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, \eta(\gamma_0, \beta)), \tag{5}$$

where

$$\eta(\gamma_0,\beta)) = \sum_{j=1}^{k+1}\alpha_j \sum_{1\leq h\leq m\leq p}\beta_{j,h}\beta_{j,m}\eta_{j,2}(\gamma_0),$$
$$\eta_{j,2}(\gamma_0) = I(f)\int_{\mathbb{R}}\Big(\frac{1}{V(x)}\Big)^2\frac{\partial T}{\partial\gamma_{j,h}}(\gamma_0,x)\frac{\partial T}{\partial\gamma_{j,m}}(\gamma_0,x)dF_j(x)<\infty,$$
$$I(f) = \int_{\mathbb{R}}\phi_f^2(x)f(x)dx<\infty.$$

We considered the central sequence $\Pi_n(\gamma_0,\beta)$ expressed as follow

$$\Pi_n(\gamma_0,\beta) = \Theta_{1n} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{\frac{1}{V(\mathbf{X}_{t-1})}\beta^\top N(\gamma_0,\mathbf{X}_{t-1})\phi_f[\varepsilon_t(\gamma_0)]\right\}. \tag{6}$$

Based on the previous results and under H_0, we can write

$$\begin{pmatrix}\Pi_n(\gamma_0,\beta)\\ \Theta_n\end{pmatrix}\xrightarrow{law}\mathcal{N}\left(\begin{pmatrix}0\\ -\dfrac{\eta(\gamma_0,\beta)}{2}\end{pmatrix},\begin{pmatrix}\eta(\gamma_0,\beta) & \eta(\gamma_0,\beta)\\ \eta(\gamma_0,\beta) & \eta(\gamma_0,\beta)\end{pmatrix}\right). \tag{7}$$

By returning to testing H_0 against $H_1^{(n)}$, we consider the following statistic of test:

$$\mathcal{T}_n(\gamma_0,\beta) = \frac{\Pi_n(\gamma_0,\beta)}{\widehat{\pi}_n(\gamma_0,\beta)} \tag{8}$$

where

- $\widehat{\pi}_n(\gamma_0,\beta) = \sqrt{\widehat{\eta}_n(\gamma_0,\beta)}$,
- $\widehat{\eta}_n(\gamma_0,\beta) = \sum_{j=1}^{k+1}\widehat{\alpha}_j\sum_{1\leq h\leq m\leq p}\hat{\beta}_{j,h}\hat{\beta}_{j,m}\hat{\eta}_{j,2}(\gamma_0)$,
- $\widehat{\eta}_{j,2}(\gamma_0) = I(f)\int_{\mathbb{R}}\frac{1}{V^2(x)}\frac{\partial T}{\partial\gamma_{j,h}}(\gamma_0,x)\frac{\partial T}{\partial\gamma_{j,m}}(\gamma_0,x)dF_j(x)$
- $\widehat{\eta}_{j,2}(\gamma_0)$ is an estimator of $\eta_{j,2}(\gamma_0)$ and $\widehat{\alpha}_j$ is an estimator of $\lim\limits_{n\longrightarrow+\infty}\dfrac{n_j(n)}{n}$.

As $n\to+\infty$, under H_0, we have

$$\mathcal{T}_n(\gamma_0,\beta)\xrightarrow{law}\mathcal{N}(0,1).$$

From (7), by the Le Cam's third lemma (proposition 4.2 in [5]), under $H_1^{(n)}$, we can write

$$\Pi_n(\gamma_0,\beta) \xrightarrow{law} \mathcal{N}(\eta(\gamma_0,\beta),\eta(\gamma_0,\beta)).$$

We show that, when $n \to +\infty$, under H_0,

$$\widehat{\pi}_n(\gamma_0,\beta) \longrightarrow \pi(\gamma_0,\beta).$$

This convergence remains true under $H_1^{(n)}$ by contiguity.
By using Le cam's third lemma, we conclude that under $H_1^{(n)}$, as $n \longrightarrow +\infty$,

$$\frac{\Pi_n(\gamma_0,\beta)}{\pi(\gamma_0,\beta)} \xrightarrow{law} \mathcal{N}(\pi(\gamma_0,\beta),1).$$

Then, under $H_1^{(n)}$, as $n \to +\infty$, we have

$$\frac{\Pi_n(\gamma_0,\beta)}{\widehat{\pi}_n(\gamma_0,\beta)} \xrightarrow{law} \mathcal{N}(\pi(\gamma_0,\beta),1).$$

We used in this section many mathematical tools in order to analyse the asymtotic behavior of our test under the null and the alternatives hypotheses.

2.3 Power of the test

To calculate the power of our test statistic , we need to derive the asymptotic cumulative distribution of $\dfrac{\Pi_n(\gamma_0,\beta)}{\widehat{\pi}_n(\gamma_0,\beta)}$.
From simple computation, one has:

$$\lim_{n\to+\infty} P\Big(\frac{\Pi_n(\gamma_0,\beta)}{\widehat{\pi}_n(\gamma_0,\beta)} > z_\alpha \big| H_1^{(n)}\Big) = 1-\Phi(z_\alpha - \pi(\gamma_0,\beta)),$$

where Φ is the cunulative distribution function of a standard Gaussian distribution and z_α is its $(1-\alpha)$-quantile, $\alpha \in [0,1]$.
By using section 4.4.3 of [5], the test based on $\mathcal{T}_n(\gamma_0,\beta)$ is locally asymptotic optimal.

3 Conclusion

The aim of the present work is a preliminary study to a new changepoint detection method that we are currently investigating as a generalisation of Ngatchou-Wandji and Ltaifa [8]. The likelihood ratio test studied here is optimal. The LAN property is an important tool to find the distribution of the test under the local alternatives.

References

[1] W Hardle, BU Park, and AB Tsybakov. Estimation of non-sharp support boundaries. *Journal of Multivariate Analysis*, 55(2):205–218, 1995.

[2] Elizabeth Million. The hadamard product. *Course Notes*, 3(6), 2007.

[3] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[4] Jib Huh. Detection of a change point based on local-likelihood. *Journal of multivariate analysis*, 101(7):1681–1700, 2010.

[5] Jean-Jacques Droesbeke and Fine Jeanne. *Inférence non paramétrique: Les statistiques de rangs*. Association pour la statistique et ses utilisations, Ed. de l'Université de Bruxelles; Ed. Paris: Ellipses, 1996.

[6] Hans-Georg Muller. Change-points in nonparametric regression analysis. *The Annals of Statistics*, pages 737–761, 1992.

[7] Xiaofeng Shao and Xianyang Zhang. Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240, 2010.

[8] Joseph Ngatchou-Wandji and Marwa Ltaifa. On detecting weak changes in the mean of charn models. *arXiv preprint arXiv:2101.08597*, 2021.

Inference of multiscale gaussian graphical model

Do Edmond Sanou [1] Christophe Ambroise [2] & Geneviève Robin [3]

Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry-Courcouronnes, France.
[1] *doedmond.sanou@univ-evry.fr* [2] – *christophe.ambroise@univ-evry.fr* [3]
genevieve.robin@cnrs.fr

Résumé. Les modèles graphiques Gaussiens sont largement utilisés pour l'analyse de données génomiques. Une difficulté majeure réside dans le nombre de variables mesurées, qui excède en général le nombre d'observations de plusieurs ordres de grandeur. Pour cette raison, une réduction du nombre de variables est souvent réalisée avant de procéder à l'inférence de graphes. Nous proposons une nouvelle méthode permettant d'inférer simultanément une structure de clustering hiérarchique, et les graphes décrivant la structure d'indépendance à chaque niveau de la hiérarchie. Cette méthode repose sur la résolution d'un problème d'optimisation convexe combinant une pénalité de type lasso graphique, avec une pénalité fused lasso. Nous résolvons ce problème à l'aide d'un algorithme reposant sur le lissage de Nesterov, et présentons des résultats initiaux de simulation sur des données réelles.
Mots-clés. Modèles graphiques Gaussiens, clustering hiérarchique, fused lasso

Abstract. Gaussian graphical models are widely used for the analysis of genomic data. A major difficulty lies in the number of measured variables, which generally exceeds the number of observations by several orders of magnitude. For this reason, a reduction of the number of variables is often performed before proceeding to graph inference. We propose a new method allowing to simultaneously infer a hierarchical clustering structure, as well as the graphs describing the structure of independence, at each level of the hierarchy. This method is based on solving a convex optimization problem combining a graphical lasso penalty, with a fused type lasso penalty. We solve this problem using an algorithm based on Nesterov's smoothing, and present initial results on real data.
Keywords. Gaussian graphical models, hierarchical clustering, fused lasso

1 Introduction

Graphical models (Lauritzen, 1996) allow to visually synthesize scientific content through modular networks, and are used in several fields ranging from social sciences to genetics and economics.
Let $\mathbf{X} = (X^1, \ldots, X^p)$ be a p-dimensional Gaussian random vector, with mean μ and covariance matrix $\mathbf{\Sigma}$.

The conditional independence structure of X is characterized by a graph $G = (V, E)$, where $V = \{1, \dots p\}$ is the set of vertices and E the set of edges, uniquely determined by the support of the precision matrix $\Omega = \Sigma^{-1}$ (Dempster, 1972). In other words, for any two vertices $i, j \in V$, the edge (i, j) belongs to the set E if and only if $\Omega_{ij} \neq 0$. In addition, $\Omega_{ij} = 0$ if and only if the i-th and j-th variables are conditionally independent given the others.
The inference of Gaussian graphical models thus boils down to inferring the support Ω, which is generally assumed to be sparse, leading to sparse graph inference (Meinshausen and Bühlmann 2006, Friedman et al. 2008).
See, e.g., Fan (2016) for a review of sparse graph inference methods. In addition to sparsity, structural assumptions on the underlying graph structure are often used during the inference task. For instance, clustering structures (Ambroise et al. 2009, Tan et al. 2013, Devijver and Gallopin 2017, Cheng et al. (2017)).
We propose a novel method to estimate simultaneously a hierarchical clustering structure, and graphical models depicting the conditional independence structure between meta-variables, at each level of the hierarchy. The procedure is based on a convex optimization problem with a hybrid penalty term combining a graphical lasso and a fused lasso penalty. In the spirit of convex hierarchical clustering, introduced by Hocking et al. (2011), the hierarchy is obtained by spanning the entire regularization path. We solve the optimization problem using an efficient convex optimization algorithm based on the continuation of Nesterov's smooting technique (Hadj-Selem et al. 2018), and present initial results on real data.

2 Multiscale Graphical Lasso

High-dimensional inference of Gaussian graphical models often relies on sparse estimation of the precision matrix. The neighborhood selection introduced by Meinshausen et Bühlmann (2006) is a well known example which estimates the network structure by regressing each variable in the data set on the others, and applying lasso regularization. Let the $n \times p$-dimensional matrix $\boldsymbol{X}$ contains n independent observations of a Gaussian p-dimensional vector. Let's denote $\hat{\beta} = (\hat{\beta}_i)_{1 \leq i \leq p} \in \mathbb{R}^{p \times p}$ the estimator of regression coefficients; Meinshausen et Bühlmann (2006) define $\hat{\beta}$ through the convex program

$$\hat{\beta} = \underset{\substack{\beta \in \mathbb{R}^{p \times p} \\ \beta_i^i = 0 \text{ for all } i}}{\arg\min} \left\{ \sum_{i=1}^{p} \frac{1}{n} ||X_i - \boldsymbol{X}\beta_i||_2^2 + \lambda ||\beta_i||_1 \right\}.$$

The neighborhood selection approach enjoys many favorable theoretical and empirical properties, and is robust to high-dimensional problems. However, in settings where the number of variables greatly exceeds the number of observations, as it is the case in genomic data analysis (number of samples $\sim$ 100 and number of variables $\sim$ 1e4), inferring clusters

of variables prior or simultaneously to graph inference is often necessary, for the sake of both statistical sanity and interpretability.
We revisit the neighborhood selection method of Meinshausen et Bühlmann (2006) with the addition of a group-fused lasso type penalty, which allows simultaneous inference of a graphical model and a hierarchical clustering of the variables.
The addition of fused type term in the inference of graphical models has been studied by authors such as Danaher et al. (2014), Ganguly et al. (2014) with a prior knowledge on the structure of the variables or observations. Meixia et al.(2020) proposed a likelihood-based method to infer also simultaneously clustering structure and a network using a l_1-norm fused norm.
We propose to minimize the following pseudo-likelihood which combines lasso and group-fused lasso penalties:

$$f(\boldsymbol{\beta}; \boldsymbol{X}) = \sum_{i=1}^{p} \frac{1}{n} ||X_i - \boldsymbol{X}\beta_i||_2^2 + \lambda_1 \sum_{i=1}^{p} ||\beta_i||_1 + \lambda_2 \sum_{i<j} ||\beta_i - \tau_{ij}(\beta_j)||_2, \tag{1}$$

where τ_{ij} is a transposition which replaces the difference $\beta_{ij} - \beta_{jj}$ by $\beta_{ij} - \beta_{ji}$. In addition to the lasso penalty term on β_i, we have therefore introduced a group-fused lasso penalty term on the pairwise differences. This penalty sums up the l_2-norm of the pairwise differences between regression vectors, forcing them to be identical as λ_2 increases. Thus, in the spirit of convex clustering introduced by Hocking et al. (2011), we get a hierarchical clustering structure by spanning the regularization path obtainted by varying λ_2 while λ_1 is fixed. Note that in the remaining of the paper, we will consider the transposition notation as implicit for the sake of simplicity.
We construct meta-variables for each cluster according to the approach defined in Park et al. (2006), which regress a response variable on the mean variables of the clusters. Thus, for each level of the clustering hierarchy, our clusters are replaced by the average of the variables that compose them. This has the advantage of reducing the number of parameters to be estimated as we go up in the hierarchy, but also of synthesizing the information by filling the clusters with representative variables. The representative variables are the average of the group variables.

3 Continuation of Nesterov's Smoothing (CONESTA)

Minimizing our convex optimization problem

$$\underset{\substack{\beta \in \mathbb{R}^{p \times p} \\ \beta_i^i = 0 \text{ for all } i}}{\arg\min} \; f(\boldsymbol{\beta}; \boldsymbol{X}), \tag{2}$$

can be casted in the framework of the CONESTA solver (Hadj-Selem et al., 2018).

CONESTA solver addresses a very general class of optimization problems including many group-wise penalties. The function that it minimizes has the form

$$f(\beta) = g(\beta) + \lambda_1 h(\beta) + \lambda_2 s(\beta),$$

where $g(\beta)$ is a least square criterion with ridge loss, $h(\beta)$ is a penalty whose proximal is know (for example the ℓ_1 loss) and $s(\beta)$ is a $\ell_{1,2}$ loss of the form

$$s(\beta) = \sum_{\phi} \|A_\phi \beta_\phi\|_2,$$

where A_ϕ are linear operators and β_ϕ subvectors of $vec(\beta)$. CONESTA then approaches the $l_{1,2}$-norm with a smooth function which gradient is known after applying a Nesterov smoothing. The smoothing parameter is then updated dynamically with respect to the distance to the minimum of the cost function. As this minimum is not known, CONESTA proposes an upper bound of the distance to the minimum using the duality gap. A FISTA (Beck and teboulle, 2009) algorithm is then used to solve the optimization problem for a fixed value of the smoothing parameter.
Our graphical inference problem can be reformulated for the CONESTA solver by specifying the linear operators A_Φ for the $\ell_{1,2}$ group norm:

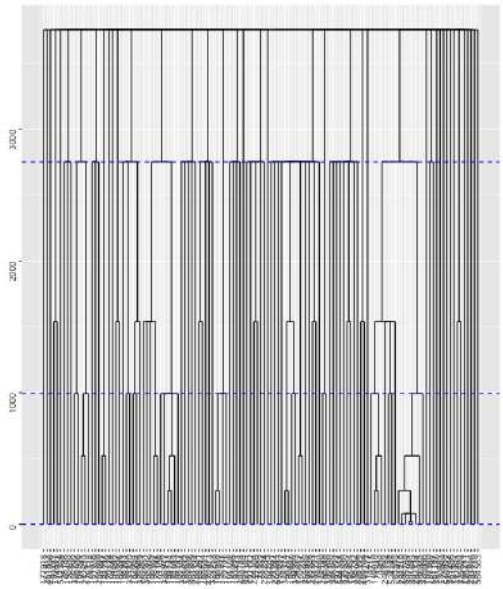
$$s(\beta) = \sum_{\phi} \|A_\phi \beta_\phi\|_2 = \sum_{i<j} ||\beta_i - \beta_j||_2 = \sum_{i<j} ||A_{ij}\text{vec}(\beta)||_2 \tag{3}$$

where A_{ij} is a $p \times p^2$ matrix defined by

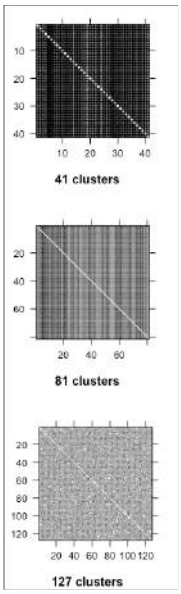
$$A_{ij}(k,\ell) = \begin{cases} 1, \text{ if } \ell = (i-1)p + k, \\ -1, \text{ if } \ell = (j-1)p + k, \\ 0, \text{ otherwise.} \end{cases}$$

4 Numerical illustration

We illustrate our new method for inference of multiscale Gaussian graphical model, with an application to the analysis of microbial associations from the American Gut project. The dataset used is a subset composed of $n = 25$ samples and $p = 127$ microbial operational taxonomic units counts. We normalized the data by applying the centered log-ratio normalization as proposed by Kurtz et al. (2015) and also by Rau (2017).
We fix starting values of l_1-norm and l_{12}-norm penalties at $1e^{-4}$. They are increased iteratively as we get higher in the hierarchy. The threshold for OTUs fusion, that is euclidean distance between regression vectors, is fixed to $1e^{-5}$.
The visualization of estimated graphs is proposed for three levels in the hierarchy in Figure 1. For each level, an associated network is availble. This allows as we go up in the hierarchy to compress the information.



(a) Dendrogram.



(b) Cluster assignments and independence graphs for $k = 127$ (graphical lasso) $k = 81$ and $k = 41$ clusters.

Figure 1: Variable fusions and graphical models for three different compression scales.

Bibliography

Ambroise, C., Chiquet, J., Matias, C. (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3(0), 205–238.

Beck, A., Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm. *Society for Industrial and Applied Mathematics Journal on Imaging Sciences*, 2(1), 183–202.

Cheng, L., Shan, L., Kim, I. (2017). Multilevel Gaussian graphical model for multilevel networks. *Journal of Statistical Planning and Inference*, 190, 1–14.

Danaher, P., Wang, P., Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes.

Dempster, A. P. (1972). Covariance Selection. *Biometrics*, 28(1), 157.

Devijver, E., Gallopin, M. (2018). Block-Diagonal Covariance Selection for High-Dimensional Gaussian Graphical Models. *Journal of the American Statistical Association*, 113(521), 306–314.

Fan, J., Liao, Y., Liu, H. (2016). An overview of the estimation of large covariance and precision matrices.

Friedman, J., Hastie, T., Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso.

Ganguly, A., Polonik, W., Associate, P. (2014). Local Neighborhood Fusion in Locally Constant Gaussian Graphical Models.
Hadj-selem, F., Lofstedt, T., Dohmatob, E., Frouin, V., Dubois, M., Guillemot, V., Duchesnay (2018). Continuation of Nesterov ' s Smoothing for Regression with Structured Sparsity in High-Dimensional
Hocking, T. D., Joulin, A., Bach, F., Vert, J.-P. (2011). Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties.
Johnson RW (1996), Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*
Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology, 11(5), e1004226.*
Lauritzen, S. L. (1996). Graphical models. Clarendon Press.
Meinshausen, N., Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3), 1436–1462.
Park, M. Y., Hastie, T., Tibshirani, R. (2006). Averaged gene expressions for regression. Biostatistics, 8(2), 212–227.
Rau, A. (2017). Statistical methods and software for the analysis of transcriptomic data.

Reconstruction de la connectivité fonctionnelle en Neurosciences: une amélioration des algorithmes actuels

Gilles Scarella [1] & Cyrille Mascart [2] & Alexandre Muzy [3] & Tien Cuong Phi [4] & Patricia Reynaud-Bouret [5]

[1] *Université Côte d'Azur, CNRS, LJAD/I3S - gilles.scarella@univ-cotedazur.fr*
[2] *Université Côte d'Azur, CNRS, I3S - mascart@i3s.unice.fr*
[3] *Université Côte d'Azur, CNRS, I3S - alexandre.muzy@cnrs.fr*
[4] *Université Côte d'Azur, CNRS, LJAD - Tien_cuong.phi@univ-cotedazur.fr*
[5] *Université Côte d'Azur, CNRS, LJAD - reynaudb@univ-cotedazur.fr*

Résumé. Afin d'identifier la connectivité fonctionnelle entre neurones, des travaux précédents (dans [5] notamment) ont utilisé des processus de Hawkes pour modéliser les intensités conditionnelles des trains de spikes et ont reconstruit la connectivité fonctionnelle par moindres carrés pénalisés. On propose ici une nouvelle méthode de construction des matrices du même problème Lasso obtenu et l'utilisation d'un autre solveur, qui semble plus efficace, pour une amélioration du temps de calcul.

Mots-clés. Neurosciences, processus de Hawkes, Lasso, connectivité fonctionnelle, méthode d'active set

Abstract. To identify the functional connectivity between neurons, previous works (see [5] in particular) have used Hawkes processes to model conditional intensities of spike trains and have reconstructed functional connectivity by penalized least square method. We propose here a new method to construct the matrices in the same resulting Lasso problem and the choice of another solver, which seems to be more efficient, to improve computational times.

Keywords. Neuroscience, Hawkes processes, Lasso, functional connectivity, active set method

1 Présentation du problème

Le but est d'étudier la connectivité fonctionnelle entre neurones, qui est un enjeu important en Neurosciences. La présente étude est basée sur l'enregistrement simultané des temps d'émission des potentiels d'action, ou spikes, de M neurones. Comme cela a été présenté dans [5], on considère M trains de spikes simultanés, modélisés comme des

processus de Hawkes multivariés. L'intensité du i-ième train de spikes N^i a la forme suivante

$$\lambda_i(t) = \left(\nu_i + \sum_{j=1}^{M} \sum_{T \in N^j, T<t} h_{j\to i}(t-T) \right)_+, \quad \forall\, t, \quad \forall i \in [\![1, M]\!].$$

Le coefficient ν_i est le taux de décharge spontané du i-ième train de spikes, donnant la fréquence moyenne d'apparition d'un nouveau spike en l'absence d'excitation ou d'inhibition, et la fonction $h_{j\to i}$, qui dépend du temps, modélise l'interaction excitatrice ou inhibitrice du j-ième train sur le i-ième. On suppose que les fonctions $h_{j\to i}$ sont constantes par morceaux sur une partition de K intervalles de taille δ, telles que:

$$h_{j\to i} = \sum_{k=1}^{K} a^k_{j\to i}\, \varphi_k \quad \text{où } \varphi_k = \mathbb{1}_{((k-1)\delta, k\delta]}.$$

Les coefficients (ν_i) et $(a^k_{j\to i})$ doivent être estimés, pour tous $(i,\, j) \in [\![1, M]\!]^2$ et $k \in [\![1, K]\!]$.

Le code *neuro-stat*[1], inclus dans le package R *UnitEvents*[2], introduit dans [5], permet de retrouver une estimation *sparse* des coefficients (ν_i) et $(a^k_{j\to i})$ et donc du graphe de connectivité fonctionnelle (où l'existence d'une connexion entre j et i correspond à la non nullité de $h_{j\to i}$). De plus, le code permet aussi, en fonction de différentes phases de l'enregistrement, d'estimer différents jeux de paramètres et différents graphes (typiquement on peut alors associer un graphe à un comportement animal).

Ici nous nous intéressons à l'amélioration de l'algorithme sur une plage de temps fixée $(T_{\min}, T^{\max}]$.
De manière générale, on se focalise sur le critère des moindres carrés suivant, que nous pénaliserons ensuite:

$$\int_{T_{\min}}^{T^{\max}} \overline{\lambda_i}^2(t)\, dt - 2 \int_{T_{\min}}^{T^{\max}} \overline{\lambda_i}(t)\, dN^i_t$$

où $\overline{\lambda_i}$ est un candidat intensité de la forme

$$\begin{aligned} \overline{\lambda_i}(t) &= \overline{\nu_i} + \sum_{j=1}^{M} \sum_{T \in N^j, T<t} \overline{h}_{j\to i}(t-T), \quad \forall\, t, \quad \forall i \in [\![1, M]\!], \\ \text{avec } \overline{h}_{j\to i} &= \sum_{k=1}^{K} \overline{a}^k_{j\to i}\, \varphi_k \end{aligned}$$

Soit $\psi^l_t(\varphi_k)$ la fonction prévisible définie par:

$$\psi^l_t(\varphi_k) = \int_{-\infty}^{t^-} \varphi_k(t-u)\, dN^l_u = \sum_{T<t, T\in N^l} \mathbb{1}_{((k-1)\delta, k\delta]}(t-T), \tag{1}$$

[1] `https://github.com/ybouret/neuro-stat`
[2] `https://sourcesup.renater.fr/frs/?group_id=3267`

on en déduit de (1) que $\overline{\lambda}_i$ s'écrit donc sur le dictionnaire des fonctions prévisibles sous la forme suivante

$$\overline{\lambda}_i(t) \;=\; \overline{\nu}_i + \sum_{j=1}^{M}\sum_{k=1}^{K} \overline{a}^k_{j\to i}\,\psi^j_t(\varphi_k)$$

En introduisant comme dans [5] les matrices b et G, de dimension $(1+MK)$-by-M (resp. $(1+MK)$-by-$(1+MK)$), et en notant $\alpha(l,k) = 1+(l-1)\,K+k$, l'indice dans $[\![2, 1+M\,K]\!]$, pour $k \in [\![1,K]\!]$ et $l \in [\![1,M]\!]$, on voit que le critère des moindres carrés devient

$$-2\,{}^T\mathbf{b}^i\,\beta + {}^T\beta\,\mathbf{G}\,\beta \quad \forall i \text{ avec}$$

$$\begin{aligned}
\mathbf{b}^i_{\alpha(l,k)} &= \int_{T_{\min}}^{T^{\max}} \psi^l_t(\varphi_k)\,dN^i_t \quad \text{et} \qquad \mathbf{b}^i_1 = \#\left\{T\in N^i,\;\; T\in(T_{\min},T^{\max}]\right\},\\
\mathbf{G}_{\alpha(l_1,k_1),\alpha(l_2,k_2)} &= \int_{T_{\min}}^{T^{\max}} \psi^{l_1}_t(\varphi_{k_1})\,\psi^{l_2}_t(\varphi_{k_2})dt, \quad \mathbf{G}_{\alpha(l,k),1} = \int_{T_{\min}}^{T^{\max}} \psi^l_t(\varphi_k)\,dt \text{ et } \mathbf{G}_{1,1}=T^{\max}-T_{\min}
\end{aligned} \tag{2}$$

En suivant [3], on pénalise par une norme l_1 à poids tels que $\mathbf{d}$ est une matrice de dimension $(1+MK)$-by-M définie à partir de $\mu_{\mathbf{2}}$ (de même dimension) et $\mu_{\mathbf{A}}$ vecteur de taille $(1+MK)$.

On utilise la même définition que [5] pour $\mathbf{d}$, dans laquelle on prend $\gamma = 3$.

$$\begin{aligned}
\mathbf{d}^i &= \sqrt{2\gamma\,c_{log}\,\mu_{\mathbf{2}i}} + \frac{\gamma}{3} c_{log}\,\mu_{\mathbf{A}},\;\; \forall i\in[\![1,1+MK]\!],\;\; c_{log} = \log((1+MK)M)\\
(\mu_{\mathbf{A}})_{\alpha(l,k)} &= \sup_{t\in(T_{\min},T^{\max}]} |\psi^l_t(\varphi_k)|,\;\; (\mu_{\mathbf{A}})_1 = 1,\\
(\mu_{\mathbf{2}})^i_{\alpha(l,k)} &= \int_{T_{\min}}^{T^{\max}} \left(\psi^l_t(\varphi_k)\right)^2\,dN^i_t,\;\; (\mu_{\mathbf{2}})^i_1 = \#\left\{T\in N^i,\;\; T\in(T_{\min},T^{\max}]\right\}.
\end{aligned} \tag{3}$$

On suit la même procédure que dans [5] et l'on doit résoudre un ensemble de problèmes Lasso de dimension $(1+M\,K)$, pour tout $i\in[\![1,M]\!]$, pour obtenir les coefficients $a^k_{j\to i}$

$$\begin{aligned}
\mathbf{a}^i_{BL} = \arg \min_{\beta\in\mathbb{R}^{(1+MK)}} & -2\,{}^T\mathbf{b}^i\,\beta + {}^T\beta\,\mathbf{G}\,\beta + 2\,{}^T\mathbf{d}^i\,|\beta|\\
\text{avec} \quad & \mathbf{a}^i_{BL} = (\nu_i, a^1_{1\to i}, a^2_{1\to i}, \cdots, a^K_{1\to i}, a^1_{2\to i}, \cdots, a^K_{M\to i})\\
\text{et} \quad & |\beta| \;= (|\beta_1|, |\beta_2|, \cdots, |\beta_{MK}|, |\beta_{1+MK}|)
\end{aligned} \tag{4}$$

2 Nouvelle méthode et résultats

La méthode mise en œuvre dans [5], dans le code *neuro-stat*, est coûteuse en temps calcul et utilisable uniquement dans le logiciel R, ce qui conduit à certaines limitations. En effet, cette méthode utilisait des calculs d'intégrales de fonctions constantes par morceaux pour définir les matrices intervenant dans le problème d'estimation. Si $Ntot$ désigne le nombre total de spikes, la complexité était alors en $O(M\,Ntot\,K^2)$ pour $\mathbf{G}$ et en $O(M\,Ntot\,K)$ pour $(\mathbf{b}, \mu_{\mathbf{A}}, \mu_{\mathbf{2}})$.

L'objectif est de considérer ici entre 1 000 et 10 000 neurones, soit un nombre supérieur à celui considéré dans [5]. La référence est le *BlueBrain project*[3] qui simule des colonnes corticales d'environ 10 000 neurones.

La nouvelle méthode présentée ici utilise, comme la précédente, des fonctions à support borné mais est moins coûteuse: on utilise la portée $A = K\delta$ de telle sorte que l'influence du spike τ_1 est nulle sur le spike τ_2 si $|\tau_1 - \tau_2| > A$. Dans la nouvelle méthode de calcul, T est un tableau de taille $Ntot$, contenant les valeurs des temps de spikes dans l'ordre croissant et indépendamment des neurones; $neur$ est un tableau de même taille que T contenant le numéro de neurone du spike correspondant. La Figure 1 illustre la nouvelle méthode utilisée pour calculer $\mathbf{G}$, $\mathbf{b}$, $\mu_{\mathbf{2}}$ et $\mu_{\mathbf{A}}$, où l'on passe en revue chacun des spikes et l'on affecte sa contribution à l'indice approprié de la quantité à calculer.

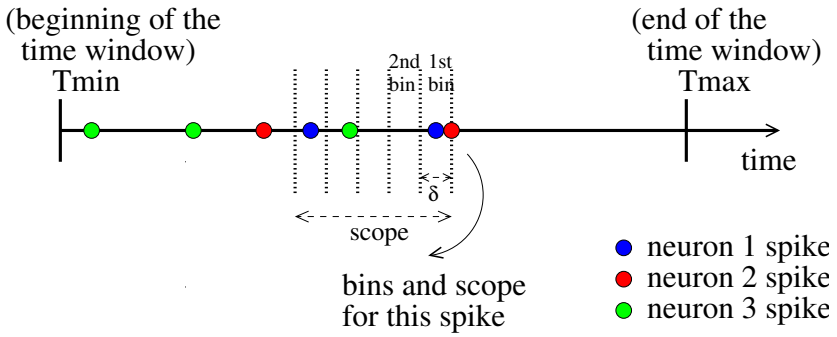


Figure 1: Exemple pour M=3 neurones et K=5 bins

Par exemple, on obtient pour les différents termes de $\mathbf{G}$ définie dans (2):
Pour tous $l \in [\![1, M]\!]$, $k \in [\![1, K]\!]$, on a

$$\mathbf{G}_{1,\alpha(l,k)} = \sum_{\substack{\theta \in N^l, \\ (T_{\min} - A) \leqslant \theta < T^{\max}}} \left(\min\left(T^{\max}, k\,\delta + \theta\right) - \max\left(T_{\min}, (k-1)\,\delta + \theta\right)\right),$$

[3] `https://www.epfl.ch/research/domains/bluebrain/`

On obtient pour tous $(k_1, k_2) \in [\![1, K]\!]^2$, $(l_1, l_2) \in [\![1, M]\!]$

$$\mathbf{G}_{\alpha(l_1,k_1),\alpha(l_2,k_2)} = \sum_{\substack{(\theta,\tau)\in(N^{l_1}\times N^{l_2}),\\ (T_{\min}-A\leqslant\theta<\tau<T^{\max})}} (\min(\theta+k_1\delta,\ \tau+k_2\delta,\ T^{\max}) - \max(\theta+(k_1-1)\delta,\ \tau+(k_2-1)\delta,\ T_{\min}))$$

Le calcul de $\mathbf{G}$ est décrit en Figure 2 (avec une numérotation des indices commençant à 0).

Algorithm 1 Computation of matrix G

```
1: function COMPUTE_G(DataSpike, T_min, T_max)
2:     i_begin = INDEX(T_min − Kδ, DataSpike)
3:     i_end = INDEX(T_max, DataSpike)-1 ⟶ [-1 for getting the largest spike strictly smaller than T_max]
4:     g[0, 0] = T_max − T_min
5:     for θ ∈i_begin:i_end do
6:         for k ∈ 1 : K do
7:             dx = min(T_max,T[θ]+kδ)-max(T_min, T[θ]+(k-1)δ)
8:             if dx > 0 then
9:                 g[0, neur[θ]K + k] += dx
10:                g[neur[θ]K + k, 0] += dx
11:        for k_1 ∈ 1 : K do
12:            x_1=min(T_max, T[θ]+k_1δ)
13:            dx = x_1-max(T_min, T[θ]+(k_1-1)δ)
14:            if dx > 0 then
15:                g[neur[θ]K + k_1, neur[θ]K + k_1] += dx
16:            for k_2 ∈ (k_1 + 1) : K do
17:                dx = x_1-max(T_min, T[θ]+(k_2-1)δ)
18:                if dx > 0 then
19:                    g[neur[θ]K + k_1, neur[θ]K + k_2] += dx
20:                    g[neur[θ]K + k_2, neur[θ]K + k_1] += dx
21:        for τ ∈ (θ + 1):end and T[τ]<T[θ]+Kδ do
22:            for k_1 ∈ 1 : K do
23:                for k_2 ∈ 1 : k_1 do
24:                    dx = min(T_max, T[θ]+k_1δ, T[τ]+k_2δ) - max(T_min, T[θ]+(k_1 − 1)δ, T[τ]+(k_2 − 1)δ)
25:                    if dx > 0 then
26:                        g[neur[θ]K + k_1, neur[τ]K + k_2] += dx
27:                        g[neur[τ]K + k_2, neur[θ]K + k_1] += dx
```

Figure 2: Nouvel algorithme pour G

On calcule les autres quantités définies dans (2) et (3) en suivant le même procédé.

Les Figures 3 et 4 montrent une comparaison des temps de construction de $\mathbf{G}$ entre *neuro-stat* et la nouvelle méthode, sur deux exemples, le premier à $Ntot$ fixé ($Ntot \simeq 9.0\,10^6$), simulant des processus de Poisson; le second tel que $Ntot{=}M\,\nu(T^{\max} - T_{\min})$ où ν est une fréquence donnée ($\nu = 20$ Hz), simulant des processus de Hawkes sur l'intervalle de temps $(0, 100]$ avec le code SPIKES (voir [6]). La nouvelle méthode est plus efficace.

Pour le calcul des estimateurs définis dans (4), on utilise désormais la méthode d'Active Set semblable à celle expliquée dans [1] afin de réduire les temps calculs. Pour chaque problème Lasso intermédiaire, on utilise un solveur LARS (cf [2]).

Sur la Figure 5, le test considéré ici est un processus de Hawkes utilisant le code SPIKES [6] sur (0,100] avec un taux de décharge spontanée de 10 Hz pour chaque neurone et un nombre moyen de connexions pour chaque neurone égal à 25. Pour la reconstruction, on choisit $\delta = 0.02\,s$. Après comparaison à *neuro-stat*, les valeurs numériques obtenues sont identiques entre les deux méthodes (à la précision numérique près).

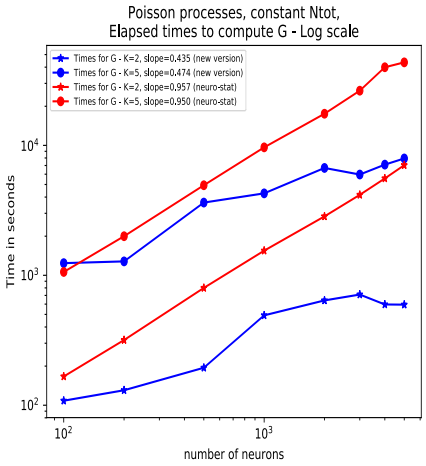


Figure 3: Comparaison des temps calcul de $\mathbf{G}$ - $K = 2$ et 5 - Test 1

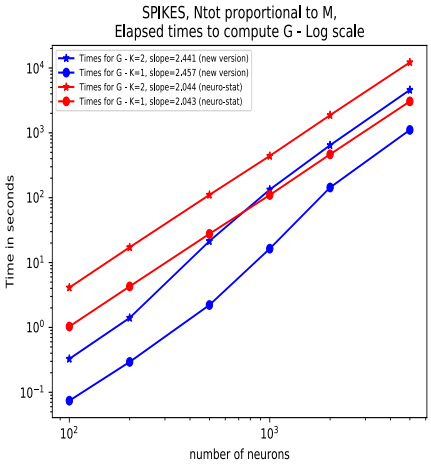


Figure 4: Comparaison des temps calcul de $\mathbf{G}$ - $K = 1$ et 2 - Test 2

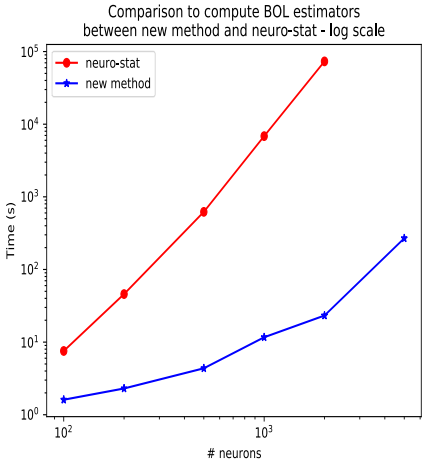


Figure 5: Comparaison des temps calcul des estimateurs - Test 3

En utilisant des trains de spikes simulés à l'aide du code SPIKES décrit dans [6], une série de tests a été menée sur des réseaux de grande taille, entre 5 000 et 20 000 neurones. Le Tableau 1 présente les caractéristiques de chaque réseau.

M	nombre de spikes	Tmin (s)	Tmax (s)	taux de décharge (Hz)	coefficients d'interaction
5 000	20 410 540	0	100	20	25
10 000	85 572 676	0	100	20	25
10 000	107 032 361	0	100	75% à 20 Hz, 12.5% à 10 Hz, 12.5% à 40 Hz	25
10 000	128 228 049	0	100	50% à 10 Hz, 50% à 50 Hz	25
15 000	121 800 711	0	100	20	25
20 000	163 450 693	0	100	20	25

Table 1: Description des réseaux

Le Tableau 2 montre les temps calcul obtenus pour les trains de spikes du Tableau 1.

test	M	temps calcul pour **G** (s)	temps pour b (s)	temps pour muA (s)	temps pour mu_2 (s)	temps pour estimateur BOL (s)	mémoire max (Go)
1	5 000	5 405.72	2 190.49	0.34	1 736.79	691.82	3.06
2	10 000	34 395.4	14 432	1.37	23 605.3	2 181.3	11
3	10 000	53 100.8	21 491.7	1.71	42 457.8	2 083.58	11
4	10 000	69 998.9	30 334.2	2.05	64 523.9	2 254.43	11
5	15 000	74 096.1	30 063.3	1.95	48 759.5	6 087.92	26
6	20 000	134 760	56 184.6	2.64	90 188.4	11 492	47

Table 2: Temps calcul pour les exemples du Tableau 1

3 Conclusion et perspectives

On a expliqué comment améliorer le temps de calcul de la matrice $\mathbf{G}$ par rapport à la méthode précédente et on a montré des résultats de comparaison. On procède de manière semblable pour améliorer les temps calcul pour $\mathbf{b}$, $\mu_{\mathbf{2}}$ et $\mu_{\mathbf{A}}$. Les temps calcul des estimateurs du problème Lasso ont aussi été réduits en utilisant la méthode d'Active Set.

Des premiers tests pour des nombres de neurones élevés ont été menés. L'article [4] est une référence récente présentant des résultats de reconstruction de la connectivité pour des réseaux allant jusqu'à 10 000 neurones, à l'aide des corrélations croisées.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2020-A0080311481 attribuée par GENCI.

Bibliographie

[1] Dragoni, L., Flamary, R., Lounici, K., and Reynaud-Bouret, P. (2019) *Large scale Lasso with windowed active set for convolutional spike sorting*, arXiv:1906.12077.

[2] Efron, B. and Hastie, T. and Johnstone, I. and Tibshirani, R. (2004), *Least angle regression*, The Annals of Statistics, 2-32; 407–499.

[3] Hansen, N-R, Reynaud-Bouret, P. and Rivoirard, V. (2015), *Lasso and probabilistic inequalities for multivariate point processes*, Bernoulli, 21(1), 83–143.

[4] Kobayashi, R., Kurita, S., Kurth, A., Kitano, K., Mizuseki, K., Diesmann, M., Richmond, B. J. and Shinomoto, S. (2019) *Reconstructing neuronal circuitry from parallel spike trains*, Nature Communications, 10 (1).

[5] Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N. and Reynaud-Bouret, P., (2018), *Reconstructing the functional connectivity of multiple spike trains using Hawkes models*, Journal of Neuroscience Methods, 297, 9–21.

[6] Mascart, C., Muzy, A. and Reynaud-Bouret, P. (2020), *Efficient Simulation of Sparse Graphs of Point Processes*, submitted to ACM.

Sur le compromis risque-équité dans le cadre général de la régression*

Evgenii Chzhen[1] & Nicolas Schreuder[2]

[1]*Institut de Mathématiques d'Orsay, Université Paris-Sud, 91405 Orsay*
evgenii.chzhen@universite-paris-saclay.fr
[2]*CREST, 5 Avenue Le Chatelier, 91120 Palaiseau*
nicolas.schreuder@ensae.fr

Résumé. Nous proposons un cadre théorique pour l'étude du problème de l'apprentissage d'une fonction réelle qui satisfasse à des critères d'équité. Ce cadre est bâti sur la notion d'amélioration relative de niveau α de la fonction de régression, que nous introduisons à partir de la théorie du transport optimal. Le cas $\alpha = 0$ correspond au problème de régression sous contrainte de "Parité Démographique", tandis que nous retrouvons le problème usuel de régression pour $\alpha = 1$. Les valeurs intermédiaires $\alpha \in (0, 1)$ nous permettent d'interpoler de manière continue entre les deux cas pré-cités pour étudier les propriétés de prédicteurs partiellement équitables. Au sein de ce cadre, nous quantifions de manière précise le coût, en terme de risque, induit par l'introduction de la contrainte d'équité.

Mots-clés. Régression, équité, transport optimal, arbitrage.

Abstract. We propose a theoretical framework for the problem of learning a real-valued function which meets fairness requirements. This framework is built upon the notion of α-relative (fairness) improvement of the regression function which we introduce using the theory of optimal transport. Setting $\alpha = 0$ corresponds to the regression problem under the Demographic Parity constraint, while $\alpha = 1$ corresponds to the classical regression problem without any constraints. For $\alpha \in (0, 1)$ the proposed framework allows to continuously interpolate between these two extreme cases and to study partially fair predictors. Within this framework we precisely quantify the cost in risk induced by the introduction of the fairness constraint.

Keywords. Regression, fairness, optimal transport, trade-off.

1 Introduction

Data driven algorithms are deployed in almost all areas of modern daily life and it becomes increasingly more important to adequately address the fundamental issue of historical

*Titre alternatif : Sur la théorie des déblais et de remblais pour l'éthique en intelligence artificielle.

§Cette note présente quelques résultats de Chzhen and Schreuder (2020). Nous renvoyons le lecteur vers cet article pour une étude plus détaillée de ce problème et pour accéder aux preuves des résultats énoncés.

biases present in the data (Barocas et al., 2019). The goal of algorithmic fairness is to bridge the gap between the statistical theory of decision making and the understanding of justice, equality, and diversity. The literature on fairness is broad and its volume increases day by day, we refer the reader to (Barocas et al., 2019, Mehrabi et al., 2019) for a general introduction on the subject and to (del Barrio et al., 2020, Oneto and Chiappa, 2020) for reviews of the most recent theoretical advances.

Basically, the mathematical definitions of fairness can be divided into two groups (Dwork et al., 2012): *individual fairness* and *group fairness*. The former notion reflects the principle that similar individuals must be treated similarly, which translates into Lipschitz type constraints on possible prediction rules. The latter defines fairness on population level via (conditional) statistical independence of a prediction from a sensitive attribute (*e.g.*, gender, ethnicity). A popular formalization of such notion is through the *Demographic Parity* constraint, initially introduced in the context of binary classification (Calders et al., 2009). Despite of some limitations (Hardt et al., 2016), the concept of Demographic Parity is natural and suitable for a range of applied problems (Köeppen et al., 2014, Zink and Rose, 2019).

In this work we study the regression problem of learning a real-valued prediction function, which complies with an approximate notion of Demographic Parity while minimizing expected squared loss.

Unlike its classification counterpart, the problem of fair regression has received far less attention in the literature. However, as argued by Agarwal et al. (2019), classifiers only provide binary decisions, while in practice final decisions are taken by humans based on predictions from the machine. In this case a continuous prediction is more informative than a binary one and justifies the need for studying fairness in the regression framework.

Notation For any univariate probability measure μ we denote by F_μ (*resp.* F_μ^{-1}) the cumulative distribution function (*resp.* the quantile function) of μ. For two random variables U and V we denote by $\mathrm{Law}(U \mid V{=}v)$ the conditional distribution of the random variable $U \mid V{=}v$ and we write $U \overset{d}{=} V$ to denote their equality in distribution. For any integer $K \geq 1$, we denote by Δ^{K-1} the probability simplex in $\mathbb{R}^K$ and we write $[K] = \{1, \dots, K\}$. For any $a, b \in \mathbb{R}$ we denote by $a \vee b$ (*resp.* $a \wedge b$) the maximum (*resp.* the minimum) between a, b. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second-order moment.

2 Problem statement

We study the regression problem when a sensitive attribute is available. The statistician observes triplets $(\boldsymbol{X}_1, S_1, Y_1), \dots, (\boldsymbol{X}_n, S_n, Y_n) \in \mathbb{R}^p \times [K] \times \mathbb{R}$, which are connected by

the following regression-type relation

$$Y_i = f^*(\boldsymbol{X}_i, S_i) + \xi_i \ , \qquad i \in [n] \ , \tag{1}$$

where $\xi_i \in \mathbb{R}$ is a centered random variable and $f^* : \mathbb{R}^p \times [K] \to \mathbb{R}$ is the regression function. Here for each $i \in [n]$, $\boldsymbol{X}_i$ is a feature vector taking values in $\mathbb{R}^p$, S_i is a sensitive attribute taking values in $[K]$, and Y_i is a real-valued dependent variable. A prediction is any measurable function of the form $f : \mathbb{R}^p \times [K] \to \mathbb{R}$. We define the risk of a prediction function f via the $\mathbb{L}_2$ distance to the regression function f^* as

$$\mathcal{R}(f) := \|f - f^*\|_2^2 := \sum_{s=1}^{K} w_s \mathbb{E}\left[(f(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S))^2 \mid S = s\right] \ , \quad \textbf{(Risk measure)}$$

where $\mathbb{E}[\ \cdot \mid S{=}s]$ is the expectation *w.r.t.* the distribution of the features $\boldsymbol{X}$ in the group $S = s$ and $\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in \Delta^{K-1}$ is a probability vector, which weights the group-wise risks.

For any $s \in [K]$ define ν_s^* as $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s)$ – the distribution of the optimal prediction inside the group $S = s$. Throughout this work we make the following assumption on those measures, which is, for instance, satisfied in linear regression with Gaussian design.

Assumption 2.1. *The measures $\{\nu_s^*\}_{s\in[K]}$ are non-atomic and have finite second moments.*

Regression with fairness constraints

Any predictor f induces a group-wise distribution of the predicted outcomes $\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s)$ for $s \in [K]$. The high-level idea of *group fairness* notions is to bound or diminish an eventual discrepancy between these distributions.

We define the *unfairness* of a predictor f as the sum of the weighted distances between $\{\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s)\}_{s\in[K]}$ and their common barycenter *w.r.t.* the Wasserstein-2 distance

$$\mathcal{U}(f) := \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s W_2^2\big(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s), \nu\big) \ . \qquad \textbf{(Unfairness measure)}$$

In particular, since the Wasserstein-2 distance is a metric on the space probability distributions with finite second-order moment $\mathcal{P}_2(\mathbb{R}^d)$, a predictor f is such that $\mathcal{U}(f) = 0$ if and only if it satisfies the Demographic Parity (DP) constraint defined as

$$\big(f(\boldsymbol{X}, S) \mid S = s\big) \overset{d}{=} \big(f(\boldsymbol{X}, S) \mid S = s'\big), \quad \forall s, s' \in [K] \ . \qquad \textbf{(DP)}$$

In the literature those predictions that do not satisfy the DP constraint often called as those that produce *Disparate Impact*. Exact DP is not necessarily desirable in practice

and it is common in the literature to consider *relaxations* of this constraint. In this work we introduce the α-*Relative Improvement* (α-RI) constraint – a novel DP relaxation based on our unfairness measure. We say that a predictor f satisfies the α-RI constraint for some $\alpha \in [0,1]$ if its unfairness is at most an α fraction of the unfairness of the regression function f^*, that is, $\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)$. Importantly, the fairness requirement is stated relatively to the unfairness of the regression function f^*, which allows to make a more informed choice of α.

Formally, for a fixed $\alpha \in [0,1]$, the goal of a statistician in our framework is to build an estimator $\hat{f}$ using data, which enjoys two guarantees (with high probability)

$$\textbf{α-RI guarantee:} \quad \mathcal{U}(\hat{f}) \leq \alpha \mathcal{U}(f^*) \qquad \text{and} \qquad \textbf{Risk guarantee:} \quad \mathcal{R}(\hat{f}) \leq r_{n,\alpha,f^*} \ .$$

The former ensures that $\hat{f}$ satisfies the α-RI constraint. In the latter guarantee we seek the sequence r_{n,α,f^*} being as small as possible in order to quantify *two effects*: the introduction of the α-RI *fairness constraint* and the *statistical estimation*.

3 Oracle α-relative improvement

The natural question that we address is: assuming that the underlying distribution of $X \mid S$ and the regression function f^* are known, which prediction rule f^*_α minimizes the expected squared loss under the α-RI constraint $\mathcal{U}(f^*_\alpha) \leq \alpha \mathcal{U}(f^*)$?

This section is devoted to the study of the α-relative improvement f^*_α on population level, that is, in this section we study

$$f^*_\alpha \in \arg\min \{\mathcal{R}(f) \,:\, \mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)\} \ , \qquad \forall \alpha \in [0,1] \ . \tag{2}$$

The characterization of f^*_0 (no fairness relaxation) was provided by (Chzhen et al., 2020, Le Gouic et al., 2020), we include their result to keep the presentation self-contained.

Theorem 3.1. *Let Assumption 2.1 be satisfied, then*

$$\min \Big\{ \mathcal{R}(f) \,:\, \big(f(\boldsymbol{X}, S) \mid S = s\big) \overset{d}{=} \big(f(\boldsymbol{X}, S) \mid S = s'\big) \ \forall s, s' \in [K] \Big\} = \mathcal{U}(f^*) \ . \tag{3}$$

Moreover, the distribution of the minimizer of the problem on the l.h.s. *is given by*

$$\arg\min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s W_2^2 \left(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s), \nu\right) \ .$$

An important consequence of Theorem 3.1 is that it puts the risk $\mathcal{R}$ and the unfairness $\mathcal{U}$ – two conflicting quantities – on the same scale. In particular, it allows to measure both fairness and risk using the same unit measurements, hence, study the trade-off between the two. However, Theorem 3.1 says nothing about the whole family of oracle α-RI $\{f^*_\alpha\}_{\alpha \in [0,1]}$ The next result establishes a closed form solution to the minimization Problem (2) under Assumption 2.1 for any value of $\alpha \in [0,1]$.

Theorem 3.2. *Let Assumption 2.1 be satisfied, then for all* $\alpha \in [0,1]$ *and all* $(\boldsymbol{x}, s) \in \mathbb{R}^p \times [K]$ *(up to a set of null measure) it holds that*

$$f^*_\alpha(\boldsymbol{x}, s) = \sqrt{\alpha} f^*_1(\boldsymbol{x}, s) + (1-\sqrt{\alpha}) f^*_0(\boldsymbol{x}, s) \ .$$

Recall that $f^* = f^*_1$, hence the α-relative improvement f^*_α is the point-wise convex combination of exactly fair prediction f^*_0 and the regression function f^*_1. Besides, setting $\alpha = 0$ we recover the result of Chzhen et al. (2020), Le Gouic et al. (2020) as a particular case of our framework. The set of oracle α-RI $\{f^*_\alpha\}_{\alpha\in[0,1]}$ satisfies the following properties.

1. **Risk and fairness monotonicity:** if $\alpha \leq \alpha'$, then $\mathcal{R}(f^*_\alpha) \geq \mathcal{R}(f^*_{\alpha'})$ and $\mathcal{U}(f^*_\alpha) \leq \mathcal{U}(f^*_{\alpha'})$.
2. **Point-wise convexity:** for all $\alpha, \alpha' \in [0,1]$ and all $\tau \in [0,1]$ it holds that $\tau f^*_\alpha + (1-\tau) f^*_{\alpha'} \in \{f^*_\alpha\}_{\alpha\in[0,1]}$. Moreover $\tau f^*_\alpha + (1-\tau) f^*_{\alpha'} = f^*_{\bar{\alpha}}$ with $\bar{\alpha} = (\tau\sqrt{\alpha} + (1-\tau)\sqrt{\alpha'})^2$.
3. **Order preservation:** for all $s \in [K], \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$, if $f^*(\boldsymbol{x}, s) \geq f^*(\boldsymbol{x}', s)$, then for all $\alpha \in [0,1]$ it holds that $f^*_\alpha(\boldsymbol{x}, s) \geq f^*_\alpha(\boldsymbol{x}', s)$.

The first property is intuitive and does not require the result of Theorem 3.2. The second property can be directly derived using the expression of f^*_α and it describes additional algebraic structure of the family $\{f^*_\alpha\}_{\alpha\in[0,1]}$. The third group-wise order preserving property of f^*_α is particularly attractive. Its proof is straightforward after the observation that $F_{\nu^*_s}$ and $\sum_{s'=1}^K w_{s'} F^{-1}_{\nu^*_{s'}}$ are non-decreasing functions and the fact that the composition of two non-decreasing functions is non-decreasing. For the special case of $\alpha = 0$, this observation has already been made in (Chzhen et al., 2020) and a practical algorithm that follows the group-wise order preservation property was proposed by Plečko and Meinshausen (2020). In words, this property says: given any two individuals $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ from the same sensitive group $s \in [K]$, if the optimal prediction $f^*(\boldsymbol{x}, s)$ for $\boldsymbol{x}$ is larger than that for $\boldsymbol{x}'$, then across all levels α of fairness parameter the oracle α-RI f^*_α is not changing this order.

3.1 Risk-fairness trade-off on the population level

The next key result of our framework establishes the risk-fairness trade-off provided by the parameter $\alpha \in [0,1]$ on the population level. In particular, it establishes a simple user-friendly relation between the risk and unfairness of α-relative improvement. Note that such a result is not available neither for $\mathcal{U}_{\mathrm{TV}}$ nor for $\mathcal{U}_{\mathrm{KS}}$, due to fundamentally different geometries of the squared risk and the aforementioned distances.

Theorem 3.3. *Let Assumption 2.1 be satisfied, then for any* $\alpha \in [0,1]$ *it holds that*

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \mathcal{R}(f^*_0) = (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \ . \tag{4}$$

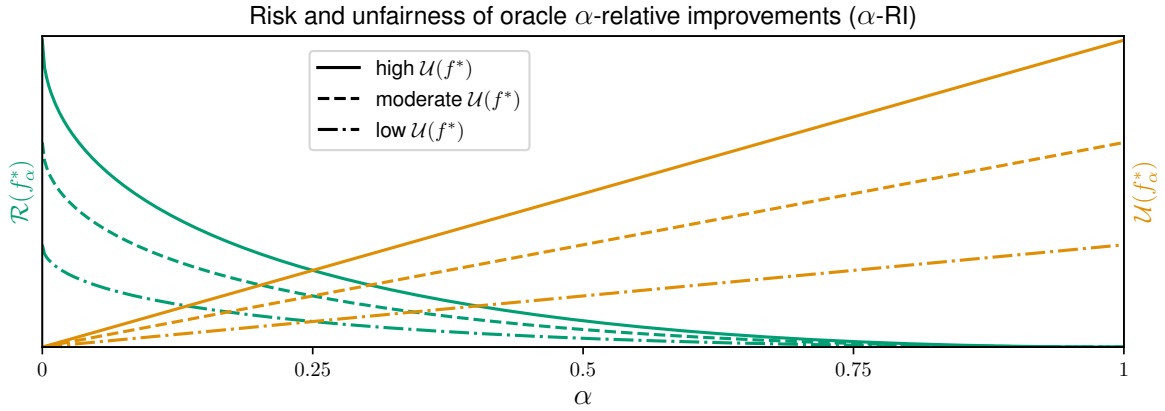


Figure 1: Risk $\mathcal{R}$ and unfairness $\mathcal{U}$ of α-RI oracles $\{f^*_\alpha\}_{\alpha\in[0,1]}$. Green curves (decreasing, convex) correspond to the risk, while orange curves (increasing, linear) correspond to the unfairness. Each pair of curves (solid, dashed, dashed dotted) corresponds to three regimes: high, moderate, and low unfairness of the regression function f^* respectively.

Observe that f^*_0, which is the optimal fair predictor in terms of DP, has the highest risk and the lowest unfairness, while the situation is reversed for $f^*_1 \equiv f^*$ – the risk is the lowest and the unfairness is the highest. Since the function $\alpha \to (1-\sqrt{\alpha})^2$ grows rapidly in the vicinity of zero, even a mild relaxation of the exact fairness constraint ($\alpha = 0$) yields a noticeable improvement in terms of the risk while having a low unfairness inflation. That is, one can find a prediction f whose unfairness $\mathcal{U}(f)$ is smaller than that of f^* by a constant multiplicative factor, without a large increase in risk. For instance, the risk of $f^*_{1/2}$ is only around 8.5% of the risk of f^*_0, while its fairness is two times better than that of f^*. This observation is illustrated in Figure 1.

4 Conclusion

This work introduces a new way to measure violation of Demographic Parity using the Wasserstein barycenter formulation. We showed that the risk-fairness trade-off in this context is completely described by one and only one parameter – the unfairness of the regression function f^*.

References

A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 2019.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009.

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*, 2020.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with Wasserstein barycenters. *NeurIPS*, 2020.

E. del Barrio, P. Gordaliza, and J.-M. Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

M. Köeppen, K. Yoshida, and K. Ohnishi. Evolving fair linear regression for the representation of human-drawn regression lines. In *2014 International Conference on Intelligent Networking and Collaborative Systems*, pages 296–303, 2014.

T. Le Gouic, J. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

L. Oneto and S. Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020.

D. Plečko and N. Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 2020.

A. Zink and S. Rose. Fair regression for health care spending. *Biometrics*, n/a(n/a), 2019.

UN SYSTEME DE CLASSEMENT PLUS JUSTE POUR LA POURSUITE EN BIATHLON

Rémi Servien

INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, remi.servien@inrae.fr

Résumé. Le biathlon est un sport olympique combinant du ski de fond et du tir, avec une pénalité pour chaque cible ratée. Les biathlètes concourent dans différents formats de course, notamment la poursuite. Durant cette course, ils poursuivent le vainqueur du sprint de la veille avec un temps de départ équivalent aux résultats du sprint. Ce format de course met en jeu des qualités comme la tactique ou la gestion de la pression émotionnelle due aux face-à-face qui ne sont pas présentes dans des courses contre-la-montre comme le sprint. Pourtant, les classements de la poursuite sont très fortement corrélés à ceux du sprint, ce qui empêche une remontée spectaculaire après un sprint raté. Nous présentons ici un nouveau classement pour la poursuite afin de pallier ce problème. Ce système de classement simple est basé sur des comparaisons avec les précédentes poursuites. Il est ensuite comparé à la version actuelle du classement des poursuites sur une course puis sur différentes saisons de coupe de monde, à partir d'une base de données de 148 poursuites masculines. Le nouveau classement modifie fortement le classement d'une poursuite mais ces modifications sont lissées à l'échelle d'une saison entière. Les avantages et limites de ce classement sont ensuite discutés, ouvrant la voie à une modification du classement des poursuites permettant de le rendre plus juste et de favoriser les surprises et le suspense dans ces courses.

Mots-clés. Biathlon, Poursuite, Système de classement.

Abstract. Biathlon is an Olympic sport combining cross-country skiing with rifle shooting, giving a penalty for each target miss. The biathletes ran different race formats, including the pursuit race. During this race, the biathletes chase the leader with a start time identical to the result of the sprint race previously achieved. So, pursuit involves different skills (such as tactics or management of emotional pressure) that are not present during races with an interval-start procedure like sprint. Nevertheless, final pursuit rankings are strongly correlated to sprint ones, which prevents a spectacular comeback after a disappointing sprint race. We present here a new pursuit ranking system to solve this issue. This simple new ranking system is based on comparisons with previous pursuit results. The current and the new rankings were then compared on a single pursuit ranking and different pursuit world cup rankings, using a database of 148 results from men pursuit world cups. The new ranking was shown to strongly modify a single pursuit ranking but these modifications were smoothed on a whole world cup season. Advantages and limitations of the new ranking system are discussed, paving the way to a fairer modification of the current pursuit ranking to increase surprise and suspense in biathlon pursuit races.

Keywords: Biathlon, Pursuit race, Ranking system.

1. Introduction

Le biathlon est un sport olympique dans lequel on combine 3 à 5 tours de ski de fond avec du tir. Entre chaque tour de ski, les biathlètes essaient d'atteindre 5 cibles situées à 50 mètres en alternant entre les positions couchée et debout. Une pénalité (en temps ou en distance supplémentaire à

parcourir) est ensuite donnée pour chaque cible manquée. Le biathlète avec le plus petit temps final remporte la course. Plusieurs formats de course existent : l'individuelle, le sprint, la poursuite et la mass-start (International Biathlon Union, 2020). Lors des poursuites, les 60 meilleurs biathlètes du sprint poursuivent le vainqueur du sprint avec un temps de départ identique à celui de l'arrivée du sprint (*i.e.* si le 2nd arrive 12s après le vainqueur lors du sprint, il partira avec 12s de retard à la poursuite). La poursuite, comme la mass-start, comprend donc des confrontations directes, où les biathlètes se battent entre eux et non pas contre le temps. Dans ces courses, la tactique joue un rôle majeur, ainsi que la gestion de ses émotions lors des ultimes sessions de tirs qui décident bien souvent du podium final (Vickers et al., 2007). Aussi, se placer de manière optimale dans un peloton est un critère essential durant ces courses (Laaksonen et al., 2018). Enfin, lors des poursuites, le temps de ski a moins d'impact sur le résultat final que lors du sprint (Laaksonen et al., 2018). On s'attend donc à ce que la poursuite ne récompense pas les mêmes qualités que le sprint.

Le biathlon n'a été que très rarement étudié de manière scientifique, mis à part pour l'impact de certains paramètres sur la précision des tirs (Josefsson et al., 2020) ou l'influence des différentes phases sur les résultats des individuelles ou des sprints (Luchsinger et al., 2019). Malgré leurs particularités, la poursuite et la mass-start demeurent quasiment non étudiées. Récemment, Luchsinger et al. (2020) ont tout de même démontré que le résultat des sprints jouait à plus de 50% dans ceux de la poursuite. Ce résultat, combiné au fait que les sprints représentent 40% des courses, implique que plus de 55% du résultat final de la coupe du monde est directement imputable au sprint. De plus, les qualités spécifiques liées à la poursuite ne sont que très rarement récompensées par les classements actuels, principalement cachées par l'importance des résultats en sprint. Un nouveau classement pour la poursuite, qui permettrait de minorer l'influence du sprint, serait donc d'un intérêt tout particulier. Différents classements ont été développées dans bien des sports (voir par exemple Kovalchik (2020) ou la revue de Wunderlich et Memmert (2020)) mais, à notre connaissance, aucun ne s'adapte facilement au contexte très spécifique de la poursuite en biathlon.

2. Méthodologie

Les résultats finaux des sprints et des poursuites sont disponibles publiquement sur la base de données de l'IBU https://biathlonresults.com/. Les résultats ont été collectés le 15 Décembre 2020 en partant de la saison 2001/2002, ce qui nous a donné 148 courses. Toutes les poursuites ont été collectées et rassemblées afin de donner la Figure 1, dans laquelle nous pouvons voir le classement d'arrivée de la poursuite en fonction de certains classements de départ, égal au classement final du sprint. Cette figure met une nouvelle fois en lumière l'importance du rang de départ pour la poursuite dans le résultat final.

Nous proposons donc de définir un nouveau classement plus juste pour la poursuite. Ce classement est basé sur un simple calcul de quantile. Pour un biathlète démarrant la poursuite au rang k, le quantile q_{ki} est calculé en fonction de la position de son rang d'arrivée f_{ki} dans la distribution des rang d'arrivée de tous les précédents biathlètes avec le rang de départ i. Certaines de ces distributions sont représentées dans la Figure 1.

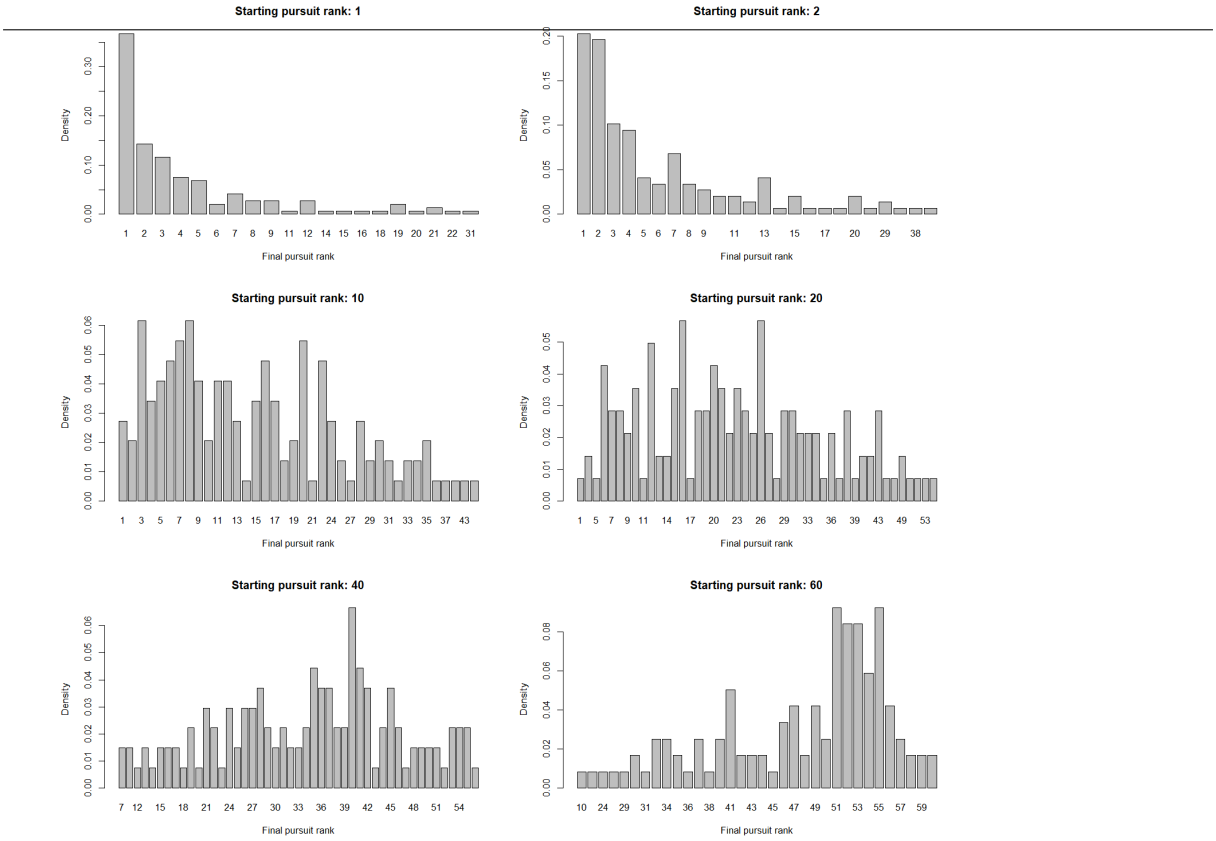


Figure 1. Diagrammes en bâtons du classement final actuel des poursuites en fonction de six différents rangs de départ.

Plus précisément, q_{ki} est donné par la formule suivante

$$q_{ki} = 1 - \frac{\sum_{j=1}^{148} \mathbb{1}_{(f_{ji} \geq f_{ki})}}{148}$$

où f_{ji} est le rang final du biathlète avec le rang de départ i à la poursuite j. Les quantités $(q_{ki})_{i=1,\ldots,60}$ sont ensuite ordonnées, ce qui permet d'obtenir le nouveau classement de la poursuite k. En cas d'égalité, le meilleur rang d'arrivée à la poursuite est privilégié, ce qui assure que le premier à franchir la ligne d'arrivée sera également le premier au nouveau classement. Cette formule est relativement naturelle : en effet, si q_{ki} est égal à zéro (resp. 1) cela signifie que, durant les 148 poursuites, aucun biathlète avec ce rang de départ n'a obtenu de meilleur (resp. pire) rang f_{ki} et que, par conséquent, ce biathlète mérite un très bon (resp. très mauvais) nouveau classement final pour la poursuite.

3. Résultats

Nous étudions tout d'abord l'influence du nouveau classement sur une poursuite spécifique, celle qui a eu lieu en 2019 à Annecy-Le Grand Bornand. Les résultats sont donnés dans le Tableau 1.

Classement actuel pour la poursuite	Nom	Classement du sprint	Nouveau classement pour la pousuite	Gain
1	BOE Johannes Thingnes	4	1	0
2	FILLON MAILLET Quentin	3	5	-3

3	CHRISTIANSEN Vetle	13	2	1
4	BOE Tarjei	2	25	-21
5	DOLL Benedikt	1	42	-37
6	JACQUELIN Emilien	20	4	2
7	FOURCADE Martin	12	12	-5
8	BJOENTEGAARD Erlend	5	24	-16
9	PEIFFER Arnd	21	10	-1
10	SCHEMPP Simon	32	3	7
12	DALE Johannes	6	38	-26
32	BORMOLINI Thomas	60	6	26

Tableau 1. Classement du sprint et nouveau et actuel classement pour la poursuite pour la course d'Annecy-Le Grand Bornand en 2019 pour certains biathlètes (pour le tableau complet voir Servien, 2021). Le gain est la différence entre le nouveau et l'actuel classement de la poursuite.

La corrélation entre le rang de départ et le rang final actuel (resp. le nouveau rang final) de la poursuite est de 0.82 (resp. 0.20), ce qui met en lumière l'influence diminuée du sprint sur le nouveau classement de la poursuite. Si nous regardons les principales modifications, nous pouvons voir que T. Boe, B. Doll, E. Bjoentegaard ou J. Dale perdent plus de 15 rangs avec le nouveau classement ce qui illustre le fait que leurs bons classements actuels sont principalement dus à leurs bons résultats en sprint. D'un autre côté, E. Jacquelin, S. Schempp, and T. Bormolini ont réalisé de belles performances durant la poursuite (resp. 14, 22, 28 rangs gagnés durant la course) et méritent leurs meilleurs classements en utilisant le nouveau système. Par exemple, T. Bormolini serait classé 6e en utilisant le nouveau classement alors que, dans les 148 poursuites précédentes, aucun biathlète partant 60e n'a fait mieux que 10e avec le classement actuel. A l'échelle d'une course, l'influence du sprint semble donc amoindrie et des remontées plus spectaculaires sont possibles.

Nous allons maintenant étudier l'influence de ce nouveau classement à l'échelle d'une saison de coupe du monde, pour les 10 dernières saisons. Tout d'abord les corrélations entre les rangs de départs et les rangs actuels sont en moyenne de 0.74 avec le classement actuel mais seulement de 0.06 avec le nouveau classement. Ensuite si on regarde les éventuelles modifications à l'échelle de la saison, il y a plus de biathlètes marquant des points (*i.e.* finissant dans les 40 premiers d'au moins une poursuite) avec le nouveau classement pour les 10 saisons qu'avec l'actuel, avec une hausse moyenne de 11 biathlètes. Au niveau du nombre de points marqués, les différences entre le 1er et les rangs de 2 à 10 sont également plus faibles en moyenne avec le nouveau classement. Ces classements plus serrés auraient donné lieu à plus de suspense lors des courses finales de la saison.

Si l'on regarde les modifications sur les podiums finaux des coupes du monde, on peut remarquer qu'elles sont plus faibles que ce qu'on aurait pu attendre en voyant les modifications créées à l'échelle d'une seule course. En effet, en comparant les classements actuels et nouveaux sur les 10 dernières saisons, nous avons 7 fois le même vainqueur, deux fois une inversion de places entre le premier et le second et la dernière fois le 4^{e} du classement actuel passant 1er avec le nouveau classement. Il y a seulement deux podiums parfaitement identiques mais, si on regarde les biathlètes sur ces podiums, 23 sur 30 sont communs aux deux classements. Ceci permet de mettre en avant une majorité de traits communs entre les deux modes de classement même si certaines situations individuelles peuvent être fortement modifiées. Comme exemple le plus marquant nous avons un biathlète qui était 3^{e} avec le classement actuel et qui se retrouve 16^{e} avec le nouveau classement, ce qui met en avant l'importance de ses bonnes performances en sprint dans ses bons classements en poursuite.

Il est à noter que ces modifications peuvent également avoir un impact majeur sur le classement général de la coupe du monde (incluant les 4 différentes épreuves). En effet, en 2019/2020, J. Boe

s'est finalement imposé au classement général avec seulement 2 points d'avance sur M. Fourcade. Avec le nouveau classement de la poursuite, ce serait M. Fourcade qui se serait imposé avec la même marge. Bien sûr, ceci est de la science-fiction car l'utilisation de ce nouveau classement aurait probablement modifié le comportement des biathlètes mais cela permet de remarquer qu'il est possible de diminuer l'impact du sprint dans le classement général de la coupe du monde (J. Boe en avait gagné 4 cette année-là) en utilisant le nouveau classement de la poursuite, notamment les années où le classement général est très serré.

Plus de détails sur les résultats sont disponibles dans Servien (2021).

4. Discussion

Le principal avantage du nouveau classement est bien entendu la moindre importance donnée aux résultats du sprint dans le résultat final. En effet, même le 60^{e} à la fin du sprint a une chance de monter sur le podium ce qui permettrait aux futures poursuites de gagner en surprise et en suspense et diminuerait l'influence démesurée du sprint dans le classement général de la coupe du monde. Ensuite, si le nouveau classement modifie en profondeur le résultat de chaque course, il modifie largement moins les résultats finaux à l'échelle d'une saison de coupe du monde de poursuite. Ceci paraît logique dans la mesure où, même si certaines qualités sont spécifiques à chaque format de course, cela reste du biathlon avec du ski de fond et du tir. Par conséquent, les meilleurs biathlètes restent les mêmes, le nouveau classement permettant de définir la poursuite comme une discipline à part entière avec des vrais spécialistes, pas seulement comme une petite perturbation des résultats du sprint.

Des critiques peuvent cependant être faites au nouveau classement présenté ici. En effet, il est plus compliqué que le classement actuel et nécessite quelques centièmes de secondes de calcul alors qu'avec le classement actuel, si on passe la ligne 3^{e} on est 3^{e}. Mais cette critique doit être atténuée. Tout d'abord, le vainqueur du nouveau classement est bien le premier qui passe la ligne et est par conséquent connu immédiatement. Ensuite, pour le sprint ou l'individuelle (ou dans d'autres sports comme le ski alpin ou le décathlon), les rangs finaux ne sont connus que quand le dernier participant coupe la ligne d'arrivée. Il est également possible, à chaque temps de passage, de calculer très vite le nouveau classement afin d'informer les biathlètes en temps réel de leur classement comme cela est fait actuellement.

Une autre limite est le fait que, quand plusieurs biathlètes ne prennent pas le départ de la poursuite ou abandonnent en cours de course, cela fait artificiellement monter les performances des biathlètes franchissant la ligne finale de la poursuite en bas de classement. Cela peut donc donner des bons nouveaux classements non mérités car principalement dus à un nombre important d'abandons et non pas à une importante remontée durant la course. Ceci peut être facilement réglé en intégrant le nombre de biathlètes finissant chaque course dans le calcul de q_{ki}. Néanmoins, cela complexifie un peu la formule et donc, afin de la garder simple et compréhensible et comme ce genre de cas est rare et n'influence que peu les meilleurs rangs du nouveau classement, cela n'a pas été pris en compte ici.

En conclusion, nous pouvons donc dire que le nouveau classement présenté ici est moins corrélé au classement du sprint que le classement actuel. Certaines limites demeurent mais, si elles sont considérées comme rédhibitoires, elles peuvent être corrigées sans difficulté. Cette communication ouvre donc la voie à un nouveau classement plus juste pour la poursuite en biathlon qui permettrait d'augmenter facilement les surprises et le suspense dans ces courses.

Bibliographie

International Biathlon Union. (2020). IBU event and competition rules. http://www.biathlonworld.com/downloads/ (accessed on 8 January 2021).

Josefsson, T., Gustafsson, H., Iversen Rostad, T., Gardner, F. L., and Ivarsson, A. (2020). Mindfulness and shooting performance in biathlon. A prospective study. *European Journal of Sport Science*, Forthcoming. doi: 10.1080/17461391.2020.1821787.

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. *International Journal of Forecasting,* 36(4), 1329-1341. doi: 10.1016/j.ijforecast.2020.01.006.

Laaksonen, M. S., Jonsson, M., and Holmberg, H.-C. (2018). The Olympic Biathlon – Recent Advances and Perspectives After Pyeongchang. *Frontiers in Physiology*, 9, 796. doi: 10.3389/fphys.2018.00796.

Luchsinger, H., Kocbach, J., Ettema, G., and Sandbakk, Ø. (2019). The Contribution From Cross-Country Skiing and Shooting Variables on Performance-Level and Sex Differences in Biathlon World Cup Individual Races. *International Journal of Sports Physiology and Performance*, 14(2), 190-195. doi: 10.1123/ijspp.2018-0134.

Luchsinger, H., Kocbach, J., Ettema, G., and Sandbakk, Ø. (2020). Contribution from cross-country skiing, start time and shooting components to the overall and isolated biathlon pursuit race performance. *PLoS ONE*, 15(9), e0239057. doi: 10.1371/journal.pone.0239057.

Servien, R. (2021). A fairer ranking system for biathlon pursuit races. *Submitted*. https://hal.inrae.fr/hal-03120424.

Vickers, J. N., and Williams, A. M. (2007). Performing under pressure: the effects of physiological arousal, cognitive anxiety, and gaze control in biathlon. *Journal of Motor Behavior*, 39(5), 381-94. doi: 10.3200/jmbr.39.5.381-394.

Wunderlich, F., and Memmert, D. (2020). Forecasting the outcomes of sports events: A review. *European Journal of Sport Science*, Forthcoming. doi: 10.1080/17461391.2020.1793002.

Marius Soltane

Laboratoire Manceau de Mathématiques, Le Mans Université.
Joint work with Alexandre Brouste, Chunhao Cai and Longmin Wang.

Détection d'une rupture dans les processus autorégressifs à bruits gaussiens dépendants.

Dans une modélisation chronologique, la détection d'une rupture au sein des coefficients gouvernant la distribution du processus est un domaine d'étude très important. Il permet de prévenir des erreurs de prévisions mais surtout d'affiner la procédure d'estimation des paramètres sur l'échantillon non concernée pas une rupture. Nous nous proposons dans cet exposé de présenter une procédure statistique pour déterminer un changement de paramètre au sein d'un processus autorégressif généré par des bruits gaussiens stationnaires. Nous expliquerons dans un premier la procédure statistique répondant à ce problème dans le cas d'un processus autorégressif classique généré pas des bruits blancs puis nous expliquerons comment généraliser ce test dans le modèle à bruits gaussiens stationnaires. Nous illustrerons au fur et à mesure de la présentation les divers résultats via des simulations.

Testing for the change of the mean-reverting parameter of an autoregressive model with stationary Gaussian noise.

The problem of testing wether or not a change has occurred in the parameter value of an autoregressive (AR) models has been considered in [4] where the noise of the AR model is supposed to be white and weakly stationary with finite fourth moment. In particular, the first order AR model has been intensively studied In diverse fields of statistical applications, scientists have observed the long memory phenomenon where correlations between observations decay slower that independent data or AR models. Long memory phenomenon appearing in macro-level economic time series or in insurance springs from aggregation of short-memory models. It could also be found throughout spatial models via partial differential equations or critical phenomena in physics

In this talk, the detection of a change in the mean-reverting parameter of an AR(p) model with stationary Gaussian noise is studied.

In the classical case of testing for a change in an AR process with white noise, invariance principle with strongly mixing condition [7] and the application of Horvàth's extension of Darling-Erdös result for the maximum of a k-dimensionnal Ornstein-Uhlenbeck process [6] are used. But in our setting, the proof relies on the linear martingale filtering [1, 3], almost-sure refinements of the estimates of the filtered process [8] and an invariance principle without strongly mixing conditions [5].

Basics for the martingale filtering of an AR(p) model with regular stationary Gaussian noise are reminded. Then, the main result concerning the convergence of the likelihood ratio test statistics to the Gumbel extreme value distribution is given. Simulations for observation samples of finite size are also done. Details of these results are shown in [2].

Key words : Darling-Erdös Theorem, Martingales, Maximum Likelidood-ratio test, Strong invariance principle.

Références

[1] Brouste A., Cai C. and Kleptsyna M. (2014) *Asymptotic properties of the MLE for the autoregressive process coefficients under stationary Gaussian noises*, Mathematical Methods of Statistics, 23(2), 103–115.

[2] Brouste A., Cai C. Soltane M. and Wang L. (2019) *Testing for the change of the mean-reverting parameter of an autoregressive model with stationary Gaussian noise*, preprint.

[3] Brouste A. and Kleptsyna M. (2012) *Kalman type filter under stationary noises*, Systems & Control Lettes 61, 1229–1234.

[4] David R., Huang D. and Yao, Y. (1995) *Testing for a change in the parameter values and order of an autoregressive model*, The Annals of Statistics, 23(1), 282–304.

[5] Eberlein E. (1986) *On strong invariance principles under dependence assumptions*, The Annals of Probability, 14, 260–270.

[6] Horvàth L. (1993) *The maximum likelihood method for testing changes in the parameters of normal observations*, The Annals of Statistics, 21, 671–680.

[7] Kuelbs J. and Philipp W. (1980) *Almost sure invariance principles for partial sums of mixing B-valued random variables*, The Annals of Probability, 8, 1003–1036.

[8] Soltane M., *Asymptotic efficiency in autoregressive processes driven by stationary Gaussian noise*, arXiv :1810.08805, preprint.

Contact :

Brouste A., Laboratoire Manceau de Mathématiques, Le Mans Université (France)
email-adress : alexandre.brouste@univ-lemans.fr

Cai C., School of Mathematics, Shanghai University of Finance and Economics (China)
email-adress : caichunhao@mail.shufe.edu.cn

Soltane M., Laboratoire Manceau de Mathématiques, Le Mans Université (France)
email-adress : marius.soltane.etu@univ-lemans.fr

Wang, L. School of Mathematical Science, Nankai University (China)
email-adress : wanglm@nankai.edu.cn

Modèles de mélange pour le partitionnement avec données manquantes informatives

Aude Sportisse [1] & Christophe Biernacki [2] & Claire Boyer [1] & Gilles Celeux [3] & Julie Josse [4] & Fabien Laporte [5] & Matthieu Marbac-Lourdelle [6]

[1] *Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, France*
[2] *Inria Lille, Université de Lille, CNRS, France*
[3] *Inria Saclay, France*
[4] *Inria Montpellier, France*
[5] *Centre de Mathématiques Appliquées, École Polytechnique, Paris, France*
[6] *CREST, Université Rennes, Ensai, France*

Résumé Pour traiter les données manquantes, les méthodes existantes s'appliquent rarement au cas de valeurs manquantes informatives de type MNAR (Missing Not At Random) qui, bien que fréquent en pratique, nécessite la prise en compte de la distribution du processus qui cause le manque dans tout traitement statistique des données. Dans un cadre d'apprentissage non-supervisé, nous étudions les modèles de mélange pour partitionner des données (quantitatives et/ou catégorielles) contenant des variables manquantes de type MNAR. Pour ce faire, nous modélisons la le mécanisme de données manquantes, à l'aide de différentes distributions, pouvant dépendre des classes sous-jacentes et/ou des valeurs des variables manquantes elles-mêmes. Nous menons une étude théorique exhaustive de l'identifiabilité des paramètres du modèle de mélange, ainsi que de ceux du processus de manque. Afin de procéder au partitionnement des données, nous proposons des algorithmes EM ou SEM pour chacun des processus de manque proposé. Nous illustrons les méthodes sur données simulées et réelles.

Mots-clés. Modèle de mélange, Données Manquantes MNAR, Identifiabilité, Algorithmes EM et Stochastique EM.

Abstract. To deal with missing data, existing methods rarely apply to the case of informative missing values of the MNAR (Missing Not At Random) type which, though frequent in practice, requires consideration of the distribution of the process that causes the missinness in any statistical analysis of the data. In an unsupervised learning framework, we study mixture models to partition (quantitative and/or categorical) data containing missing MNAR-type variables. To do so, we model the missing data mechanism, using different distributions, which may depend on the underlying classes and/or the values of the missing variables themselves. We conduct an exhaustive theoretical study

{aude.sportisse,claire.boyer}@sorbonne-universite.fr, {christophe.biernacki,gilles.celeux,julie.josse}@inria.fr
fabien.laporte@polytechnique.edu, matthieu.marbac-lourdelle@ensai.fr

of the identifiability of the parameters of the mixture model, as well as those of the missingness process. In order to partition the data, we propose EM or SEM algorithms for each of the proposed missing processes. We illustrate the methods on simulated and real data.

Keywords. Mixture models, MNAR Missing data, Identifiability, EM and SEM algorithms.

1 Introduction et contexte

Un problème récurrent en statistique est celui des données manquantes : sur un échantillon d'individus, pour lesquels un certain nombre de variables sont mesurées, plusieurs valeurs peuvent manquer.

Notons $Y \in \mathbb{R}^{n\times d}$ le jeu de données complet (seulement partiellement connu). La présence de données manquantes est encodée à l'aide d'une matrice $C = (c_{ij})_{1\leq i\leq n, 1\leq j\leq d} \in \mathbb{R}^{n\times d}$ telle que $c_{ij} = 1$ si Y_{ij} est manquant, $c_{ij} = 0$ sinon. Les valeurs des variables observées (resp. des valeurs manquantes) pour l'individu i sont regroupées dans le vecteur y_i^{obs} (resp. y_i^{mis}). En supposant un modèle de mélange pour les données Y, notre but est d'estimer une partition (inconnue) des n individus en K groupes. Cette partition peut être encodée à l'aide de la matrice $Z = (z_1|\dots|z_n)^T \in \{0,1\}^{n\times K}$ dont la i-ème ligne $z_i = (z_{i1},\dots,z_{iK})^T \in \{0,1\}^K$ est un vecteur indicateur de groupe pour le i-ème individu, avec $z_{ik} = 1$ si y_i appartient à la classe k, et $z_{ik} = 0$ sinon. Dans ce contexte de classification non-supervisée, les données manquantes ne sont pas seulement les valeurs y_i^{mis} mais également les étiquettes de la partition z_i.

En toute généralité, les données manquantes peuvent être de différents types (identifiés par [1]) :

— manquantes complètement aléatoirement (MCAR) si la probabilité qu'une observation soit manquante est la même pour toutes les observations :

$$\mathbb{P}(c_i \mid y_i, z_i; \psi) = \mathbb{P}(c_i; \psi), \qquad \forall y_i,$$

en notant ψ le paramètre du mécanisme de données manquantes ;

— manquante aléatoirement (MAR) si l'absence d'une observation dépend seulement des variables observées :

$$\mathbb{P}(c_i \mid y_i, z_i; \psi) = \mathbb{P}(c_i \mid y_i^{\text{obs}}; \psi), \qquad \forall y_i^{\text{mis}};$$

— manquante non aléatoirement (MNAR) si l'absence d'une observation dépend des valeurs des variables manquantes (et possiblement également des variables observées).

Dans la plupart des travaux, les données manquantes sont supposées M(C)AR, facilitant ainsi l'établissement de garanties théoriques : le mécanisme à l'origine des données manquantes est dit ignorable, i.e. l'étude ne requiert pas la prise en compte de la distribution $\mathbb{P}(c_i \mid y_i, z_i; \psi)$ [2] du manque.

En pratique, il est pourtant très fréquent que les données manquantes soient MNAR. Par exemple, dans un sondage, l'absence des observations sur une variable représentant le salaire peut s'expliquer par la réticence des personnes avec haut revenu à dévoiler leur salaire. Dans le cas MNAR, ignorer le mécanisme conduirait à biaiser l'étude statistique, le mécanisme est alors dit non-ignorable : la distribution $\mathbb{P}(c_i \mid y_i, z_i; \psi)$ doit être modélisée et prise en compte.

Dans la suite, nous souhaitons procéder au partitionnement, à l'aide de modèles de mélange, de données MNAR.

2 Modélisation des données

Modèle de mélange. Afin de partitionner les données, nous supposons que leur distribution est régie par un modèle de mélange [3] de K composantes (K étant inconnu). Plus formellement, les données $y_1, \ldots, y_n$ sont considérées i.i.d. générées selon la loi de densité suivante :

$$f(y_i; \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k),$$

où $\pi_k = \mathbb{P}(z_{ik} = 1)$ est la proportion du groupe étiqueté $k \in \{1, \ldots, K\}$ (avec $\sum_{k=1}^{K} \pi_k = 1$ et $\pi_k > 0$ pour tout $k \in \{1, \ldots, K\}$), $f_k(\cdot\,; \theta_k)$ est la densité des données conditionnellement à appartenir au groupe k, paramétrée par θ_k. Les paramètres du modèle de mélange sont donc $\pi = (\pi_1, \ldots, \pi_K)$ et $\theta = (\theta_1, \ldots, \theta_K)$. Les variables en jeu peuvent être continues, catégorielles ou mixte.

Nous supposons que les données peuvent contenir des données manquantes MNAR. Ce type de mécanisme est aussi bien non-ignorable pour l'estimation de la loi de mélange que pour la tâche de clustering. Dans ce cas, la modélisation de $\mathbb{P}(c_i \mid y_i, z_i; \psi)$ est nécessaire et nous fondons notre inférence sur la vraisemblance observée complète suivante

$$L(\pi, \theta, \psi; Y^{\text{obs}}, C) = \prod_{i=1}^{n} \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) \mathbb{P}(y_i, z_{ik} = 1; \pi, \theta) dy_i,$$

où $\mathcal{Y}_i^{\text{mis}} = \{\tilde{y} = (\tilde{y}_{i1}, \ldots, \tilde{y}_{id}) \in \mathcal{Y}_i : \tilde{y}^{\text{obs}} = y^{\text{obs}}\}$. En maximisant la vraisemblance, l'intérêt est double : nous pouvons estimer les paramètres de la densité du modèle de mélange et les utiliser pour en déduire un partitionnement des données.

Modèle du mécanisme de données manquantes. Pour la distribution du processus de manque $\mathbb{P}(c_i \mid y_i, z_i; \psi)$, nous déclinons plusieurs modèles à partir du modèle général suivant :

$$\mathbb{P}(c_{ij} = 1 \mid y, z_{ik} = 1; \psi) = \rho(\alpha_{kj} + \beta_{kj} y_{ij}), \tag{1}$$

avec $\psi = (\alpha, \beta)$ et ρ la fonction de répartition d'une loi continue (e.g. les fonctions sigmoïde ou de répartition de loi normale, pour des modèles logistique ou probit).

Le paramètre α_{kj} représente l'effet moyen du lien entre la présence de valeur manquante pour la variable j et l'appartenance de l'individu correspondant à la classe k (i.e. cet effet peut ne pas être le même pour toutes les variables). Le paramètre β_{kj} représente l'effet direct du manque sur la variable j, qui peut dépendre à la fois de la classe k et de la valeur de la variable elle-même.

3 Etudes théorique et numérique

Le modèle (1) a le mérite d'être général et d'inclure de nombreuses situations, mais le grand nombre de paramètres ($2Kd$) afférents et leur estimation s'avèrent difficile à appréhender.

Identifiabilité du modèle. Théoriquement, nous montrons que le modèle (1) ne permet pas l'identifiabilité des paramètres (π, θ, ψ) de la loi du mélange et du mécanisme dans le cas où les données seraient catégorielles. L'identifiabilité reste cependant possible dans le cas continu. Nous étudions précisément ces questions d'identifiabilité pour toutes les déclinaisons du modèle suivant le type de données en jeu, mais aussi suivant les fonctions de lien ρ.

Estimation. De par la complexité de la tâche, l'utilisation d'un algorithme SEM [4] devient nécessaire avec une étape SE utilisant un algorithme de Gibbs et l'introduction d'une nouvelle variable latente. Nous proposons et explicitons des algorithmes EM [5] et SEM pour traiter tous les sous-modèles de (1), pour des données continues, catégorielles ou mixtes. Nous éprouvons les méthodes numériquement en testant leur stabilité à des erreurs de modélisation et leur capacité de sélection du nombre de classes K.

Etude d'un sous-modèle. Un cas particulier du modèle général (1) peut s'écrire comme suit

$$\mathbb{P}(c_{ij} = 1 \mid y, z_{ik} = 1; \psi) = \rho(\alpha_k), \tag{2}$$

pour lequel l'effet du manque sur la variable j ne dépend qu'exclusivement du groupe sous-jacent. Théoriquement, l'identifiabilité des paramètres (π, θ, ψ) est assurée dans les cas continu, catégoriel et mixte. Numériquement, ce modèle a de très bonnes performances en pratique, même dans les cas d'un effet omis du manque dépendant de la variable j.

4 Contributions

Les principales contributions de ce travail sont consignées ici :

— Nous présentons et illustrons un inventaire pertinent de distributions pour le processus de manque MNAR dans le cadre de la classification non-supervisée reposant sur un modèle de mélange.

— Nous menons une étude exhaustive de l'identifiabilité des paramètres du modèle de mélange (π, θ) et des paramètres du processus du manque ψ, sous certaines conditions (notamment sur le type des données et des fonctions de lien régissant la distribution du mécanisme de données manquantes). C'est un véritable enjeu dans le cadre de données MNAR, car il est fréquent que les modèles conduisent à des paramètres non-identifiables [6].

— Pour chaque modèle ou sous-modèle, un algorithme EM ou SEM est proposé, implémenté, et mis à disposition pour la reproducibilité des travaux.

— Sur données synthétiques, les différents modèles sont comparés, en terme d'ICL [7] et d'ARI [8], avec (i) un partitionnement faisant une simple hypothèse de données MCAR et (ii) des méthodes de partitionnement en deux étapes après complétion des données (les données sont d'abord imputées et des algorithmes de partitionnement classiques sont lancés sur le jeu de données complété). En particulier, nous montrerons la flexibilé du modèle MNAR (2) ne modélisant que l'effet du manque dépendent de la classe k.

— Les expériences numériques incluent également un traitement sur données réelles contenant des variables manquantes : nos méthodes sont testées sur le jeu de données médicales Traumabase [9].

Références

[1] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, 1976.

[2] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[3] Geoffrey J McLachlan and Kaye E Basford. *Mixture models : Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

[4] G. Celeux and J. Diebolt. The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2 :73–82, 1985.

[5] A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39 :1–38, 1977.

[6] G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society B*, 70 :371–388, 2008.

[7] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725, 2000.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[9] Simon I Hay, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Rizwan Suliankatchi Abdulkader, Abdishakur M Abdulle, Teshome Abuka Abebo, Semaw Ferede Abera, et al. Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2016 : a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100) :1260–1344, 2017.

When OT meets MoM: Robust estimation of Wasserstein Distance

Guillaume Staerman[1] & Pierre Laforgue [2] & Pavlo Mozharovskyi[1] & Florence d'Alché-Buc[1]

[1] *LTCI, Télécom Paris, Institut Polytechnique de Paris, surname.name@telecom-paris.fr*
[2] *DSRC & Dept. of Computer Science, Università degli Studi di Milano, Italy, pierre.laforgue@unimi.it*

Abstract. Originated from Optimal Transport, the Wasserstein distance has gained importance in Machine Learning due to its appealing geometrical properties and the increasing availability of efficient approximations. It owes its recent ubiquity in generative modelling and variational inference to its ability to cope with distributions having non overlapping support. In this work, we consider the problem of estimating the Wasserstein distance between two probability distributions when observations are polluted by outliers. To that end, we investigate how to leverage a Medians of Means (MoM) approach to provide robust estimates. Exploiting the dual Kantorovitch formulation of the Wasserstein distance, we introduce and discuss novel MoM-based robust estimators whose consistency is studied under a data contamination model and for which convergence rates are provided. Beyond computational issues, the choice of the partition size, *i.e.,* the unique parameter of theses robust estimators, is investigated in numerical experiments. Furthermore, these MoM estimators make Wasserstein Generative Adversarial Network (WGAN) robust to outliers, as witnessed by an empirical study on two benchmarks CIFAR10 and Fashion MNIST.

Keywords. Optimal transport, Robust estimation, Medians-of-Means.

1 When OT meets MoM: Robust estimation of Wasserstein Distance

Computing distances between probability distributions has become a central question in numerous modern Machine Learning applications, ranging from generative modeling to clustering. Optimal Transport (OT) [1, 2] offers an appealing and insightful tool to solve this problem, building upon the Wasserstein distance. Given two probability distributions, the latter is defined in terms of the solution to the Monge-Kantorovich optimal mass transportation problem. Interestingly, it relies on a ground distance between points to build a distance between probability distributions [3]. For that reason, the Wasserstein distance stands out from the divergences usually exploited in generative

modeling, like the f-divergences [4, 5], by its ability to take into account the underlying geometry of the space, capturing the difference between probability distributions even when they have non-overlapping supports. This appealing property has been successfully exploited in Generative Adversarial Networks (GANs) [6, 7, 8], as well as in Variational Autoencoders (VAEs) [9], where the Wasserstein distance can advantageously replace an f-divergence as the loss function. Many other applications [10, 11, 12] rely on the entropic-regularized approximations introduced by [13], which has considerably alleviated the inherent computational complexity of the Wasserstein distance in the discrete case, by drawing on the Sinkhorn-Knopp algorithm. A common feature to almost all these works is that the Wasserstein distance is estimated from finite samples. While this problem has long been theoretically studied under the i.i.d. assumption [14, 15, 16], it has never been tackled through the lens of robustness to outliers, a crucial issue in Reliable Machine Learning. Indeed, data is nowadays collected at a large scale in unmastered acquisition conditions, and through a large variety of devices and platforms. The resulting datasets often present undesirable influential observations, whether they are errors or rare observations. The presence of corrupted data may heavily damage the quality of estimators, calling for dedicated methods such as JS/TV-GANs [17] in the particular case of robust shift-parameter estimation, Robust Divergences in variational inference [18], or more general tools from robust statistics [19].

The aim of this work is to propose outliers-robust estimators of the Wasserstein distance, and illustrate their application in generative modeling. To that end, we explore how to combine a Median-of-Means approach with Optimal Transport. The Median-of-Means (MoM) is a robust mean estimator firstly introduced in complexity theory during the 1980s [20, 21, 22]. Following the seminal deviation study by [23], MoM has lately witnessed a surge of interest, mainly due to its attractive sub-gaussian behavior, under the sole assumption that the underlying distribution has finite variance [24]. Originally devoted to scalar random variables, MoM has notably been extended to random vectors [25, 26, 27] and U-statistics [28, 29]. As a natural alternative to the empirical mean, MoM has become the cornerstone of several robust learning procedures in heavy-tailed situations, including bandits [30] and MoM-tournaments [31]. A more recent line of work now focuses on MoM's ability to deal with outliers. Aside from concentration results in a contaminated context [32, 33], it has yielded promising applications in robust mean embedding [34], and the more general MoM-minimization framework [35].

In this paper, we introduce and study outliers-robust estimators of the Wasserstein distance based on the MoM methodology. Our contribution is threefold:

- Focusing on the Kantorovich-Rubinstein duality [36], we present three novel MoM-based estimators, leveraging in particular Medians of U-statistics (MoU). In the realistic setting of contaminated data, we show their strong consistency, and provide non-asymptotic bounds as well.

- We propose a dedicated algorithm to compute these three estimators in practice.

Applied on a parametric family of Lipschitz functions, *e.g.* neural networks with clipped weights, it performs a MoM/MoU gradient descent algorithm. A sensitivity analysis of the unique parameter of these estimators is also provided throught numerical experiments on toy datasets.

- We robustify WGANs (w.r.t. outliers) using a MoM-based estimator as loss function. We show the benefits of this approach through convincing numerical results on two contaminated well known benchmarks: CIFAR10 and Fashion MNIST.

References

[1] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, New York, 2003.

[2] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhauser, 2015.

[3] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[4] I. Csiszàr. Eine informationstheoretische ungleichung und ihre anwendung auf den bewis der ergodizität von markhoffschen kette. *Magyer Tud. Akad. Mat. Kutato Int. Koezl*, 8:85–108, 1963.

[5] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f -divergences. *Ann. Statist.*, 37(2):876–904, 04 2009.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS 2014)*, 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017.

[9] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[10] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[11] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Mach. Learn.*, 107(12):1923–1945, 2018.

[12] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, 2018.

[13] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, 2013.

[14] Richard M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 02 1969.

[15] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics and Probability Letters*, 76:1298–1302, 07 2006.

[16] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019.

[17] Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets, 2018.

[18] Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018).*, 2018.

[19] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics (Second Edition).* John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[20] Arkadii S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons Ltd, 1983.

[21] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[22] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

[23] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

[24] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[25] Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[26] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[27] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 04 2019.

[28] Emilien Joly and Gábor Lugosi. Robust estimation of u-statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016.

[29] Pierre Laforgue, Stephan Clémençon, and Patrice Bertail. On medians of (randomized) pairwise means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

[30] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[31] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019.

[32] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[33] P. Laforgue, G. Staerman, and S. Clémençon. How robust is the median-of-means? concentration bounds in presence of outliers. arxiv.org/abs/2006.05240, 2020.

[34] Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, and Guillaume Lecué. Monk – outlier-robust mean embedding estimation by median-of-means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

[35] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

[36] Leonid V. Kantorovich and Gennady S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

Bayesian estimation of nonlinear Hawkes processes

Déborah Sulem [1] & Vincent Rivoirard [2] & Judith Rousseau [3]

[1] *Department of Statistics, University of Oxford, deborah.sulem@stats.ox.ac.uk*
[2] *CEREMADE, University Paris Dauphine, vincent.rivoirard@dauphine.fr*
[3] *Department of Statistics, University of Oxford & University Paris Dauphine, judith.rousseau@stats.ox.ac.uk*

Résumé. Les processus ponctuels ou de comptage sont utilisés pour modéliser des données de type événements temporels ou points de l'espace, comme les catastrophes climatiques, les échanges de messages entre individus ou les transactions financières. Parmi ces processus stochastiques, le modèle de Hawkes est l'un des plus populaires pour modéliser une dépendance par rapport au passé du processus. Dans ce travail, nous nous intéressons plus particulièrement aux processus de Hawkes nonlinéaires multivariés, qui permettent de caractériser des effect d'*excitation* et d'*inhibition* entre les différentes dimensions du processus. Dans un contexte general de nonlinearité lipschitzienne et dans un cadre d'estimation bayésienne non-paramétrique, nous contrôlons le taux de concentration de la distribution a posteriori sur les paramètres, sous des hypothèses standards sur la distribution a priori et le modèle. De ces résultats nous déduisons aussi le taux de convergence d'estimateurs bayésiens.

Mots-clés. Processus de Hawkes nonlinéaires, estimation bayésienne non-paramétrique, taux de concentration.

Abstract. Point processes are counting processes that are widely applied to model event-type data such as natural disasters, online message exchanges or financial transactions. The Hawkes model is a very popular point process model in which the probability of occurrences of new events depend on the past of the process. Originally introduced to model the occurrences of earthquakes, it is notably used in neuroscience for spike trains modelling and in genetics for modelling the location of regulatory elements in the DNA. In this work we consider general nonlinear multivariate Hawkes processes, that can in particular account for *excitation* and *inhibition* phenomena between the dimensions of the process. In the context of Lipshitz nonlinearity and in a nonparametric Bayesian estimation framework, we characterize the concentration rate of the posterior distribution on the parameters, under mild assumptions on the prior distribution and the model. These results also lead to convergence rates of Bayesian estimators.

Keywords. Nonlinear Hawkes processes, nonparametric Bayesian estimation, concentration rates.

1 Nonlinear Hawkes processes

A temporal univariate point process is a counting process $(N_t)_{t\geq 0}$, where N_t denotes the number of events that have occurred until time t. It can be alternatively described by a sequence of event times $(T_1, T_2, \ldots)$ on $\mathbb{R}^+$. The probability of events' occurrences is characterized by a conditional intensity function $(\lambda_t)_t$, which is informally the infinitesimal rate of event, i.e

$$\lambda_t dt = \mathbb{P}[N_t \text{ has a jump in } [t, t+dt] | \mathcal{G}_{t^-}],$$

where $\mathcal{G}_{t^-}$ is the history of the process up to t. In the self-exciting (linear) Hawkes model, the intensity has the following form

$$\lambda_t = \nu + \int_{t-A}^{t^-} h(t-s) dN_s = \nu + \sum_{t-A \leq T_i < t} h(t - T_i).$$

The parameter $\nu > 0$ denotes the *background* - or *spontaneous* - rate of events. The function $h : \mathbb{R} \to \mathbb{R}^+$ is called the *self-exciting* function and is assumed to be integrable and to have a bounded support in $[0, A]$ where $A > 0$ is the *memory length* of the process. In this simple model, past events can only increase the probability of future occurrences (*excitation* phenomenon) and lead to the so-called *clustering effect* of events. This process is also related to continuous Galton-Watson trees [1].

More generally, in multivariate point processes, each dimension represents an entity, a location or a type of event - it is equivalent to a *marked* point process with finite mark space. In the nonlinear Hawkes model, inter-dependencies between the process' dimensions can be *excitating* and *inhibiting*. For $K \in \mathbb{N}\backslash\{0\}$, let $N = (N_t)_t = (N_t^1, \ldots, N_t^K)_t$ be a Hawkes process, where each component N_t^k records the number of events that have occurred until time t at location k. The conditional intensity $(\lambda_t)_t = (\lambda_t^1, \ldots, \lambda_t^K)_t$ is a non-linear functional of some parameters $f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K)$:

$$\lambda_t^k(f) = \phi_k \left(\nu_k + \sum_{l=1}^K \int_{t-A}^{t^-} h_{lk}(t-s) dN_s^l \right), \quad k \in [K]. \tag{1}$$

For $(l, k) \in [K]^2$, the influence of component N^l onto component N^k is parametrized by the *interaction function* $h_{lk} : \mathbb{R} \to \mathbb{R}$. It can be decomposed into an *excitating* contribution - i.e. its positive part $h_{lk}^+ = \max(h_{lk}, 0)$ - and its *inhibiting* contribution - i.e. its negative part $h_{lk}^- = -\min(h_{lk}, 0)$. Finally, the *link* function $\phi_k : \mathbb{R} \to \mathbb{R}^+$ is a non-negative and non-decreasing function.

More formally, we consider a probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ and $(\mathcal{G}_t)_t$ and $\mathcal{G}$ such that $\mathcal{G}_t = \mathcal{G}_0 \vee \sigma(N_s, s \leq t)$ with $\mathcal{G}_t \subset \mathcal{G}$ and $\mathcal{G}_0 \subset \mathcal{G}$. $(N_t)_t$ is a Hawkes process with parameter $f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K)$ if

i. almost surely, for all k, l, $(N_t^k)_t$ and $(N_t^l)_t$ never jump simultaneously;

ii. for all k the intensity process $(\lambda_t^k(f))_t$ of $(N_t^k)_t$ is given by (1).

Under certain assumptions on the parameters f and the link functions ϕ_k's, there exists almost surely a unique non-explosive pathwise process [2]. The process is also a renewal process and the regeneration times have exponential moments [3].

In this work, we intend to cover a large range of nonlinear Hawkes models, by considering Lipschitz link functions of the form :

$$\phi_k(x) = \theta_k + \psi(x), \tag{2}$$

with $\theta_k \geq 0$ and $\psi : \mathbb{R} \to \mathbb{R}^+$. We typically require that $\psi(.)$ is L-Lipschitz with $L > 0$ and $\forall x \in \mathbb{R}, \psi(x) \leq a + b\max(x, 0), a \geq 0, b \in [0, 1]$. Such nonlinear Hawkes models have been notably applied on spike-train data and are able to model resting states called *refractory periods*[1]. The parameters θ_k's can be seen as small baseline spiking rates for neurons. Several choices of functions ψ can be found in the literature [4], and include, along to their shifted and scaled variants,

- the Rectified Linear Unit (ReLU): $\psi_1(x) = \max(x, 0)$,
- a clipped exponential: $\psi_2(x) = \min(e^x, \Lambda), \Lambda > 0$,
- the sigmoid function: $\psi_3(x) = \frac{e^x}{1+e^x}$,
- the softplus: $\psi_4(x) = \log(1 + e^x)$.

2 Bayesian estimation framework

We assume that we observe a stationary Hawkes process with true parameters $f_0 = ((\nu_k^0)_{k=1}^K, (h_{lk}^0)_{k,l=1}^K)$, $\theta_0 = (\theta_k^0)_{k=1}^K$ on $[-A, T]$ with $T > 0$ and $\sigma(N_s, s < 0) \subset \mathcal{G}_0$. We consider two types of estimation problems, namely estimating the parameters f_0 and estimating the link parameters $\theta_0 = (\theta_k^0)_{k\in[K]}$ - given that the general form of the link functions ϕ_k's is known.

- *Scenario 1*: the parameter θ_0 is known and we aim at estimating f_0, with mild conditions on ψ.
- *Scenario 2*: the parameter θ_0 is unknown, and we aim at estimating (f_0, θ_0). This scenario is particularly challenging and our work only covers the case of the standard nonlinear model $\psi_1(x) = \max(x, 0)$.

To guarantee stationarity, we assume that the spectral norm of the matrix $S^0 = (L\|h_{lk}^0\|_1)_{l,k=1}^K$ is strictly less than 1, where we denote $\|h\|_1 = \int_{-\infty}^{\infty} |h(x)|dx$. The following conditions on f_0 and the link function are sufficient to identify the parameters.

Assumption 2.1

1. *There exists $I \subset \mathbb{R}_+$ a compact subset such that for any $k \in [K]$, $[\nu_k^0 - \max_l h_{lk}^{0-}{}_\infty, \nu_k^0 + \max_l h_{lk}^{0+}{}_\infty] \subset I$ and ψ^{-1} is Lipschitz on I.*

[1] A refractory period is a time interval during which a neuron is unlikely to be activated

2. *If there exists* $x_* \in \mathbb{R}$ *such that* $\psi(x_*) = 0$, *then for any* $(l,k) \in [K]^2$, $\|h_{lk}^{0-}\|_\infty < \nu_k^0 - x_*$.
3. *If there exist* $\Lambda \in \mathbb{R}$ *and* $M \in \mathbb{R}$ *such that* $\forall x \geq M, \psi(x) = \Lambda$, *then for any* $(l,k) \in [K]^2$, $\nu_k^0 + \|h_{lk}^{0+}\|_\infty < \Lambda$.

Assumption 2.2

1. $\forall \epsilon > 0, \exists M > 0, \forall x \leq M, \psi(x) \leq \epsilon$.
2. *For any* $k \in [K]$, *there exists* $l \in [K], \|h_{lk}^{0-}\|_\infty > 0$ *and there exists* $x_1 < x_2$ *and* $c_\star > 0$ *such that* $\forall x \in [x_1, x_2]$, $h_{lk}^{0-}(x) \geq c_\star$.

While in Scenario 1, only Assumption 2.1 is used for the estimation of f_0, in Scenario 2, Assumption 2.2 will guarantee that both f_0 and θ_0 can be estimated from the observations.

We denote $\mathbb{P}_0$ the stationary distribution of N, $\mathbb{P}_0(.|\mathcal{G}_0)$ its conditional distribution given $\mathcal{G}_0$ and $\mathbb{E}_0$ the expectation associated with $\mathbb{P}_0$. For $f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K)$ and $\theta = (\theta_k)_{k=1}^K$, the log-likelihood of the observation $(N_t)_{t\in[0,T]}$ has the following expression.

$$L_T(f) := \sum_{k=1}^K \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt \right],$$

The parameter spaces in our estimation scenarios are either $\mathcal{F}$ or $\mathcal{F} \times \Theta$ defined as:

$$\begin{aligned}
\mathcal{H} &= \{(h_{lk})_{l,k=1}^K;\ \|h_{lk}\|_\infty < \infty, \text{support}(h_{lk}) \subset [0,A], \forall k,l \leq K, \|S\| < 1\};\\
\mathcal{F} &= \{f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K);\ 0 < \nu_k < \infty, \forall k \leq K, (h_{lk})_{lk} \in \mathcal{H}\};\\
\Theta &= \mathbb{R}_+^K,
\end{aligned}$$

with $\|S\|$ the spectral norm of $S = (L\|h_{lk}\|_1)_{l,k=1}^K$, and the $\mathbb{L}_1$-metric:

$$\begin{aligned}
\|f - f'\|_1 &= \sum_{k=1}^K |\nu_k - \nu_k'| + \sum_{k=1}^K \sum_{l=1}^K \|h_{lk} - h_{lk}'\|_1, \quad f, f' \in \mathcal{F},\\
\|\theta - \theta'\|_1 &= \sum_{k=1}^K |\theta_k - \theta_k'|, \quad \theta, \theta' \in \Theta.
\end{aligned}$$

In Scenario 1, we consider a prior distribution Π on $\mathcal{F}$ and, as in Donnet et al. [5], the (pseudo)-posterior distribution for $F \subset \mathcal{F}$

$$\Pi(F|N) = \frac{\int_F \exp(L_T(f)) d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f)}.$$

In Scenario 2, the prior and posterior distributions are defined on $\mathcal{F} \times \Theta$, with the adequate modifications in the previous expression.

3 Main results

Our main theorems characterize the asymptotic properties of the posterior distribution $\Pi(.|N)$ under mild assumptions on the prior distribution. Informally, given a sequence ϵ_T going to 0 and a "good enough" prior distribution, the posterior distribution concentrates on balls of radius ϵ_T wrt the $\mathbb{L}_1$-distance around the true parameters f_0 (resp. (f_0, θ_0) in Scenario 2).

We define

$$B(\epsilon_T) = \{f \in \mathcal{F};\ \nu_k^0 + \epsilon_T \geq \nu_k \geq \nu_k^0, h_{lk}^0 + \epsilon_T \geq h_{lk} \geq h_{lk}^0, \forall k, l \in [K]^2\}.$$

In Scenario 2, we also define $\bar{B}(\epsilon_T) = \{(f, \theta) \in B(\epsilon_T, B) \times \Theta;\ \max_k |\theta_k - \theta_k^0| \leq \epsilon_T\}$.

Theorem 3.1 *Let N be a Hawkes process observed on $[-A, T]$ with parameters $f_0 = ((\nu_k^0)_{k=1}^K, (h_{lk}^0)_{k,l=1}^K)$, link function ψ L-Lipschitz and $\theta_0 = (\theta_k^0)_{k=1}^K$ verifying Assumption 2.1. Let $\epsilon_T = o(1)$ be a positive sequence such that $\log^3 T = O(T\epsilon_T^2)$.*

1. *If θ_0 is known (*Scenario 1*), let Π be a prior distribution on $\mathcal{F}$ satisfying the following conditions for T large enough.*

   (A0) *There exist $c_1 > 0$ such that $\Pi(B(\epsilon_T)) \geq \exp^{-c_1 T\epsilon_T^2}$.*

   (A1) *There exist subsets $\mathcal{H}_T \subset \mathcal{H}$ such that*

   $$\Pi(\mathcal{H}_T^c) = o(e^{-(\kappa_T + c_1)T\epsilon_T^2}),$$

   with $\kappa_T = O(\log^2 T)$.

   (A2) *There exist $\zeta_0 > 0$ and $x_0 > 0$ such that $\log \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, ||.||_1) \leq x_0 T\epsilon_T^2$.*

   If for some $k \in [K]$, $\theta_k^0 = 0$, we further assume either that ψ is positive and $\sqrt{\psi}$ and $\log \psi$ are Lipschitz on $\mathbb{R}^-$ or either that

   $$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right) < +\infty. \tag{3}$$

   Then for any $M_T \to +\infty$,

   $$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M_T\epsilon_T|N)\right] = o(1).$$

2. *If θ_0 is unknown and $\psi(.) = \max(., 0)$ (*Scenario 2*), let Π be a prior distribution on $\mathcal{F} \times \Theta$. We assume that Assumptions 2.2, (A0), (A1), (A2) where $B(\epsilon_T)$ is replaced by $\bar{B}(\epsilon_T)$ are satisfied and (3) holds. Then for any $M_T \to +\infty$,*

   $$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 + \|\theta - \theta_0\|_1 > M_T\epsilon_T|N)\right] = o(1).$$

Assumptions (A0), (A1), (A2) are standard assumptions of the general framework of Ghosal & van der Vaart [6]. (A0) gives a lower bound on the prior mass on balls centered at the true parameter; (A1) allows to consider growing subsets of the parameter space called sieves that concentrates most of the prior mass; (A2) bounds the complexity of those sieves. In the case where $\theta_k^0 = 0$ for some $k \in [K]$, we need additional assumptions on the true model. In particular, for $\psi(.) = \max(., 0)$, (3) is a condition on the conditional intensity of the true model which cannot be trivially expressed by general conditions on the parameters.

Finally, we deduce the convergence rate of the posterior mean estimators

$$\hat{f} = \mathbb{E}^{\Pi}[f|N] = \int_{\mathcal{F}} f d\Pi(f|N) \quad \text{(Scenario 1)},$$

$$(\hat{f}, \hat{\theta}) = \mathbb{E}^{\Pi}[(f,\theta)|N] = \int_{\mathcal{F}\times\Theta} (f \otimes \theta) d\Pi((f,\theta)|N) \quad \text{(Scenario 2)}.$$

Corollary 3.2 *Under the assumptions of Theorem 3.1, if* $\int_{\mathcal{F}} \|f\|_1 d\Pi(f) < +\infty$ *(resp.* $\int_{\mathcal{F}\times\Theta} (\|f\|_1 + \|\theta\|_1) d\Pi((f,\theta)) < +\infty$ *in Scenario 2), then for any* $M_T \to +\infty$,

$$\mathbb{P}_0 \left[\|\hat{f} - f_0\|_1 > M_T \epsilon_T \right] = o(1).$$

In addition, in Scenario 2, $\mathbb{P}_0 \left[\|\hat{\theta} - \theta_0\|_1 > M_T \epsilon_T \right] = o(1)$.

References

[1] Patricia Reynaud-Bouret and Emmanuel Roy. Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, 13(5):883 – 896, 2007.

[2] Pierre Bremaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, 1996.

[3] Manon Costa, Carl Graham, Laurence Marsalle, and Viet Chi Tran. Renewal in hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3):879–915, 2020.

[4] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[5] Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48(5):2698 – 2727, 2020.

[6] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500 – 531, 2000.

Assessing the impact of covariates in simplicial regression models

Christine Thomas-Agnan [1]

[1] Toulouse School of Economics (Université Toulouse 1 Capitole), 1, Esplanade de l'Université, 31080 Toulouse Cedex 06, FRANCE, Christine.Thomas@tse-fr.eu

Résumé. Nous considérons le problème de l'évaluation des impacts d'une variable explicative donnée dans un modèle de régression comportant des variables compositionnelles, c'est-à-dire des vecteurs à composantes positives contenant une information relative. De tels modèles sont fréquents dans beaucoup d'applications (géochimie, microbiome). Nous présentons de nouveaux types d'interprétation et les comparons avec des interprétations plus classiques. Elles sont basés sur la théorie des dérivées de fonctions d'un simplex vers un simplex et conduisent à des calculs d'élasticités et de semi-élasticités.

Mots-clés. variables compositionnelles, regression simpliciale, semi-élasticités, elasticités.

Abstract. We consider the question of evaluating the impact of a given explanatory variable in a regression model involving compositional variables, i.e. vectors with positive components carrying a relative information. Such models are frequent in many applications (geochemistry, microbiome). We present new types of interpretations and compare them with more classical ones.They are based on the theory of simplicial derivatives and lead to the computation of elasticities and semi-elasticities.

Keywords. compositional variables, simplicial regression, semi-elasticities, elasticities.

1 Introduction

Simplicial regression models are linear regression models containing compositional vector variables on the right hand side, or left hand side or on both sides of the regression equation. Compositional vectors are vectors with positive components for which we want to assess the role of the relative components (ratio of components) and possibly separately the role of the total volume (sum of the components). These type of variable are very frequent in applications: market shares in marketing, vote shares in political analysis, concentrations of different elements in geochemistry, time-use patterns in epidemiology, microorganisms composition in microbiome studies. Compositional vectors are usually associated to their representant in the unit simplex space $\mathcal{S}^{\mathcal{D}}$ ($D \in \mathbb{N}$) of positive vectors whose components add up to one. Assessing the impact of explanatory variables in such

a model is not straightforward since changing one component in such a vector cannot be done "all things equal" because all other components are also affected.

Wang et al. (2013) and Morais et al. (2018) consider this question when both dependent and independent variables are of a compositional nature. Müller et al. (2016) propose a different approach when the compositional variable is on a single side. Trinh et al. (2018) and Nguyen et al. (2021) present applications to nutrition and political science when the dependent is compositional. Coenders and Pawlowsky-Glahn (2020) and Trinh et al. (2020) focus on the case when the explanatory is compositional. Finally, Morais and Thomas-Agnan (2020) develop an interpretation based on elasticities in the general framework including the case when the total is also involved in the equation. This paper first summarizes these point of views and then goes beyond by proposing alternatives.

2 Models and notations

For a vector $\check{Z}$ of volumes (amounts) in $\mathbb{R}^{+D}$, the corresponding vector of shares (parts) in $\mathcal{S}^{\mathcal{D}}$ is obtained by closure is $Z = \mathcal{C}(\check{Z}) = (z_1/T, \cdots, z_D/T)$ where $T = \sum_{i=1}^{D} z_i$. The simplex $\mathcal{S}^{\mathcal{D}}$ is equipped with usual operations $\oplus$ and $\odot$ and the Aitchison geometry, see e.g. Pawlowsky-Glahn and Buccianti (2011). We denote by $\boldsymbol{Y}$ the dependent variable, $\boldsymbol{X}$ any compositional explanatory variable and Z any non-compositional explanatory variable.The parameters of the models are denoted $\boldsymbol{b}$ for vectors and $\boldsymbol{B}$ for matrices. We distinguish between the following two types of models

1. the dependent Y is compositional

$$\boldsymbol{Y}_i = \boldsymbol{b_0} \bigoplus_{q=1}^{Q} \boldsymbol{B}_q \boxdot \boldsymbol{X}_{qi} \bigoplus_{k=1}^{K} \boldsymbol{b}_k \odot Z_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, ..., n \tag{1}$$

2. the dependent Y is not compositional

$$Y_i = a + \sum_{k=1}^{D_X - 1} b_k^* X_{i,k}^* + \epsilon_t = a + < \boldsymbol{b}, \boldsymbol{X}_i >_A + \epsilon_i,$$

   where $\boldsymbol{X}_i, \boldsymbol{b} \in \mathcal{S}^{\mathcal{D}}$, $<,>_A$ is Aitchison inner-product and $\boldsymbol{X}_i^*$ is a transformation of $\boldsymbol{X}_i$ into coordinate space by an ilr, alr, or clr transformation (see Pawlowsky-Glahn et al. (2015) for example).

3 Elasticities

Elasticities are natural tools in some regression models involving logarithm transformations and they are widely used in econometrics. For two variables X and Y, in a log-linear

model of the type $\log Y = a + bX + \epsilon$, the semi-elasticity $\frac{\partial \mathbb{E} \log Y}{\partial X} = b$ corresponds to the parameter b and measures the relative increase of Y after an increase of one unit of X. Similarly, in a log-linear model of the type $Y = a + b \log X + \epsilon$ the semi-elasticity $\frac{\partial \mathbb{E} Y}{\partial \log X} = b$ corresponds to the parameter b and measures the absolute increase of Y after a relative increase of one per cent of X. We will show that (semi-) elasticities are also natural tools in simplicial regression models because they are linked to simplicial derivatives. Let us first consider a linear path in the simplex. Let $(\mathbf{e_m})_{m=1}^D$ is the canonical basis of $\mathcal{S}^D$ with $\mathbf{e_m} = \mathcal{C}(1, \cdots, e, \cdots 1)$. For $\delta \in \mathbb{R}$, let $\delta \odot \mathbf{e_m}$ be a simplex-increment and $\mathbf{Z} = \mathcal{C}(\check{Z}_1, \cdots, \check{Z}_D) \in \mathcal{S}_D$ an initial vector of volumes and corresponding shares. Then

$$\mathbf{Z}(h) = \mathbf{Z} \oplus \delta \odot \mathbf{e_m} = \mathcal{C}(Z_1 \exp 0, \cdots, Z_m \exp \delta, \cdots, Z_D \exp 0)$$

Hence for small δ

- if $l \neq m$, $\frac{Z_l(\delta) - Z_l}{Z_l} \sim -\delta Z_m$
- if $l = m$, $\frac{Z_m(\delta) - Z_m}{Z_m} \sim \delta(1 - Z_m)$

Therefore adding $h \odot \mathbf{e_m}$ results in increasing share Z_m by $(1 - Z_m)\delta$ per cent while decreasing the other shares by δZ_m per cent.

For a simplex-valued function $f : \mathbb{R}_+ \longrightarrow \mathcal{S}^D$ mapping $\check{x} \mapsto \mathbf{y}$, Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011) show that

$$\frac{\partial^{\oplus} f(\check{x})}{\partial \check{x}} = \lim_{\delta \to 0} \frac{1}{\delta} \odot (f(\check{x} + \delta) \ominus f(\check{x})) = \mathcal{C}\left(\exp\left(\frac{\partial \log f(\check{x})}{\partial \check{x}}\right)\right)'$$

Similarly, for a function of a simplex variable $f : \mathbb{R}_+^D \longrightarrow \mathbb{R}^k$, mapping $\check{\mathbf{x}} \mapsto y$, Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011) show that

$$\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus} x} = (\lim_{\delta \to 0} \frac{1}{\delta} (\underline{f}(\mathbf{x} \oplus \delta \odot \mathbf{e_m}) - \underline{f}(\mathbf{x})))_{m=1}^D = \frac{\partial f(\check{\mathbf{x}})}{\partial \log(\check{x})}$$

Finally, to a function $\underline{f} : \mathcal{S}^{D_1} \longrightarrow \mathcal{S}^{D_2}$, from the simplex $\mathcal{S}^{D_1}$ to the simplex $\mathcal{S}^{D_2}$, mapping $\mathbf{x} \mapsto y$ there corresponds a function $f : \mathbb{R}^{D_1} + \longrightarrow \mathcal{S}^{D_2}$, mapping $\check{\mathbf{x}} \mapsto y$, such that $\underline{f}(\mathbf{x}) = f(\check{\mathbf{x}})$ with $\mathbf{x} = \mathcal{C}(\check{\mathbf{x}})$. Then Morais and Thomas-Agnan (2020) show that

$$\begin{aligned} \frac{\partial^{\oplus} \underline{f}(\mathbf{x})}{\partial^{\oplus} x} &= \lim_{\delta \to 0} \frac{1}{\delta} (\underline{f}(\mathbf{x} \oplus \delta \odot \mathbf{e}_m) - \underline{f}(\mathbf{x})) \\ &= \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbf{f}(\check{\mathbf{x}})}{\partial \log \check{x}}\right)\right)' \end{aligned}$$

Morais and Thomas-Agnan (2020) also give further Taylor approximations linking derivatives in the simplex with elasticities and semi-elasticities. They also provide formulas for

expressing these as a function of model parameters, obtained by applying the above results to the function mapping the explanatory to the conditional expectation of the dependent, when conditinal expectation is understood in a sense adapted to simplex valued random variables.

4 Relationship between elasticities interpretations and others

We first prove that, when only Y is compositional, the semi-elasticity interpretation is very similar to the classical odds ratio interpretation in logistic regression, the relative increase of odds ratio being replaced here by the relative increase of a ratio of shares, constant throughout observations. We demonstrate a strong relationship with the interpretation in Müller et al. (2016) using pivot coordinates and considering finite increments rather than infinitesimal ones. We also compare with the interpretation proposed in Wang et al. (2013) in the case of both dependent and explanatory compositional which yields a share-ratio changes ("relative elasticities") interpretation. The advantages of our method is that we work directly in the simplex and there is no need to make a reference to a particular coordinate system. Moreover we get three possible interpretations

1. the simultaneous change in the whole set of shares
2. the relative dominance of one component with respect to the geometric mean of the others
3. the relative change of share-ratios

For the case of a compositional explanatory variable, we compare with Coenders and Pawlowsky-Glahn (2020) who discuss several reparametrizations and their corresponding interpretation. Their results are coherent with ours. They differ in the fact that they consider finite increments. We have a generic treatment for all parametrizations and we can consider more general changes in the simplex.

Finally, we consider models including a total variable and models involving spatial dependence in which we are able to derive elastiticities computations.

5 Conclusion

We demonstrate that the elasticity/semi-elasticity approach to explanatory variable impact evaluation is a natural tool in simplicial regression models and is easy to implement. We explicit its relationships with former interpretations found in the literature.

References

Coenders, G. and Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT*, 44(1).

Morais, J. and Thomas-Agnan, C. (2020). Covariates impacts in compositional models and simplicial derivatives. *Austrian Journal of Statistics*. To appear.

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47(5):1–25.

Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., and Vancakova, J. (2016). Time budget analysis using logratio methods. *arXiv preprint arXiv:1609.07887*.

Nguyen, T. H. A., Laurent, T., Thomas-Agnan, C., and Ruiz-Gazen, A. (2021). Analyzing the impacts of socio-economic factors on French departmental elections with CoDa methods. *to appear in Journal of Applied Statistics*.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons, Chichester (UK).

Trinh, H. T., Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Relations between socio-economic factors and nutritional diet in vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical methods in medical research*, page 0962280218770223.

Trinh, T.-H., Christine, T.-A., Micel, S., and Ty, B. (2020). Macronutrient balances and body mass index: New insight using compositional data analysis with a total at various quantile orders. *preprint*.

Wang, H., Shangguan, L., Wu, J., and Guan, R. (2013). Multiple linear regression modeling for compositional data. *Neurocomputing*, 122:490–500.

ShapKit: a Python module dedicated to local explanation of machine learning models

Vincent Thouvenot [1,2] & Simon Grah [3]

[1] *THALES SIX & GTS France, vincent.thouvenot@thalesgroup.com*
[2] *SINCLAIR Lab, France*
[3] *OCTO Technology, France, simon.grah.pro@gmail.com*

Résumé. Le Machine Learning connaît un succès croissant dans de nombreuses applications : défense, cybersécurité, etc. Cependant, les modèles sont souvent très complexes. Cela est problématique, en particulier pour les systèmes critiques, car les utilisateurs finaux doivent comprendre les décisions d'un algorithme (par exemple, pourquoi une alerte a été déclenchée). Une solution consiste à proposer une interprétation pour chaque prédiction individuelle en fonction de la contribution des attributs. Les valeurs de Shapley, issues de la théorie des jeux coopératifs, permettent de distribuer équitablement les contributions pour chaque attribut afin de comprendre la différence entre une valeur prédite pour une observation et une valeur de base (par exemple la prédiction moyenne d'une population de référence). Si ces valeurs présentent de nombreux avantages, elles présentent un inconvénient majeur : la complexité pour les calculer augmente exponentiellement avec le nombre de variables. Dans cet exposé, nous présentons ShapKit, un module Python développé par Thales et disponible en Open Source. Nous appliquons ShapKit sur un cas d'utilisation de la cybersécurité.

Mots-clés. Explication locale, Interprétabilité, Valeurs de Shapley.

Abstract. Machine Learning is enjoying an increasing success in many applications: defense, cyber security, etc. However, models are often very complex. This is problematic, especially for critical systems, because end-users need to fully understand the decisions of an algorithm (e.g. why an alert has been triggered or why a person has a high probability of cancer recurrence). One solution is to offer an interpretation for each individual prediction based on attribute relevance. Shapley Values, coming from cooperative game theory, allow to distribute fairly contributions for each attribute in order to understand the difference between a predicted value for an observation and a base value (e.g. the average prediction of a reference population). While these values have many advantages, including their theoretical guarantees, they have a strong drawback: the complexity increases exponentially with the number of features. In this talk, we will present and demonstrate ShapKit, a Python module developed by Thales and available in Open Source dedicated to Shapley Values computation in an efficient way for local explanation of machine learning model. We will apply ShapKit on a cybersecurity use case.

Keywords. Interpetability, Local explanation, Shapley Values.

1 Introduction

Machine Learning models are used for various applications with already successful results. Unfortunately, a common criticism is the lack of transparency associated with these algorithm decisions. This is mainly due to a greater interest in performance (measurable on specific tasks) at the expense of a complete understanding of the model. This results in a lack of knowledge of the internal working of the algorithm by the developer and the end user. The most obvious consequences are a difficulty to correct the algorithm by an expert and limiting its adoption by operational staff. Moreover, the European Commission has imposed by legal means, with the General Data Protection Regulation, this transparency constraint on companies whose algorithms learn from personal data coming from European citizens. The challenge that companies are facing today is to bring Artificial Intelligence into production. The transition from conclusive laboratory tests to a production environment is not easy. To ensure that the model generalizes well on new data, a good human/machine interaction is highly appreciated. There is no single definition of interpretability or explainability concerning model prediction (e.g. see the excellent introductory book [1]). We can separate methods into two dimensions [2]. If the method is local or global, and if its approach is model agnostic or on the contrary inherent to it. A global method aims at explaining the general behaviour of a model, whereas a local method focuses on each decision of a model. The agnostic category (also called post-hoc explanation) considers the model as a black box. On the other hand, inherent or non-agnostic methods can modify the structure of a model or the learning process to create intrinsically transparent algorithms. Naturally, the best strategy is to find a model that is both completely transparent by design and sufficient in terms of accuracy. Unfortunately, the most effective machine learning models tend to be less transparent because their degrees of complexity are high.

In this talk, we focus on local explanation and model-agnostic approaches. Moreover, we search some contrastive explanation, that means we want to compare our instance with a sub-population of instances. Finally, we assume that we work on tabular data with meaninful features. We present ShapKit, a Python module dedicated to this task and illustrate its use on a cyber security use case.

2 Shapley Values as local additive explanation: a technical perspective

Local explanation focus on a single instance and examine what the model predicts for this input, and explain why. If the Machine Learning is globally too complex to be well approximated by simplest models, locally the prediction might only depend linearly or monotonically on some features. This the idea behind LIME [3]. This approach has a major drawback: we do not guarantee the efficiency in the sense that we do not insure

that the prediction is equal to the sum of the contributions of each feature nor some others nice properties. In Collaborative Game Theory, Shapley Values [4] allows to distribute a reward fairly among players according to their contribution to the win in a cooperative. The Shapley Value of a player j is a fair share of the global wealth $v(\mathcal{M})$ produced by all players together:

$$\phi_j(\mathcal{M}, v) = \sum_{S \subset \mathcal{M} \backslash \{j\}} \frac{(d - |S| - 1)!|S|!}{d!} \left(v(S \cup \{j\}) - v(S)\right), \tag{1}$$

with $|S| = \text{cardinal}(S)$, i.e. the number of players in coalition S, $\mathcal{M}$ a set of players of dimension d, $v : P(\mathcal{M}) \to R_v$ such as $v(\emptyset) = 0$. The range R_v can be $\Re$ or a subset of $\Re$ and if $S \subset \mathcal{M}$, $v(S)$ is the amount of wealth produced by coalition S when they cooperate. The Shapley Values are the only indices which respect the four following properties :

- Additivity: $\phi_j(\mathcal{M}, v + w) = \phi_j(\mathcal{M}, v) + \phi_j(\mathcal{M}, w)$ for all j, with $v : P(\mathcal{M}) \to R_v$ and $w : P(\mathcal{M}) \to R_v$;
- Null player: if $v(S \cup \{j\}) = v(S)$ for all $S \subset \mathcal{M} \backslash \{j\}$ then $\phi_j(\mathcal{M}, v) = 0$;
- Symmetry: $\phi_{\pi j}(\pi\mathcal{M}, \pi v) = \phi_j(\mathcal{M}, v)$ for every permutation π on $\mathcal{M}$;
- Efficiency: $\sum_{j \in \mathcal{M}} \phi_j(\mathcal{M}, v) = v(\mathcal{M})$.

Shapley Values offer a solution of local explanation from additive feature importance measure class ensuring desirable theoretical properties (see e.g. [5], [6]). A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. The objective is to fairly distribute the payout around all features to obtain the prediction. We can make the following correspondence between Game Theory and model interpretability:

- The features are the players;
- The model $\hat{f}$ is the game;
- The feature attribution is the gain attribution.

A major challenge for Shapley Values computing is the number of calculus to perform: the potential coalition, and so this number, grows exponentially in function of the number of features. Some approaches to approximate them has been proposed. In this talk, we only focus on generic approximation. [6] and [7] propose a Monte Carlo approximation. Theoretic properties can be find in [8], using some concentration inequalities. The objective of [9] is to reduce the number of times the costly reward function is asked and propose an optimized version of the algorithm proposed by [6] which divides by two the use of the reward function v. Moreover, this strategy can be combined with stratified sampling

(e.g. [10], [8]) in order to reduce the number of iterations required. An alternative way of computation of Shapley Values consists to use the equivalence between Equation (1) and the Weighted Least Square (WLS) problem (see e.g. [5], [11], [9]). [9] gives the theoretical properties in term of approximation errors for this algorithm. Moreover, they propose a method based on an optimization trick of the Monte Carlo estimator. The implementation of these algorithms is proposed in *shapkit* Python module [1].

3 ShapKit overview

shapkit is a Python module to use Shapley Values as local explanation of Machine Learning model. The method is a post-hoc explanation, so the user does not have to change her usual pipeline. Firstly, the user trains as usual the model and then defines her reward function fc (e.g. simply set by the model output):

```
fc = lambda x: model.predict_proba(x)
```

The user selects an instance of interest x for which she need more interpretation and picks also one or several reference(s) (instance or dataset of individuals). If the number of features is not too high (said lower than 10), then the exact Shapley Values can be computed.

```
true_shap = ShapleyValues(x=x, fc=fc, ref=reference)
```

If the dimension is too high, the user can use some approximation approaches:

- Monte Carlo algorithm:

  ```
  mc_shap = MonteCarloShapley(x=x, fc=fc, ref=reference, n_iter=1000)
  ```

- Batch Monte Carlo algorithm: If the reward function fc can handle data set of inputs, the batch version of Monte Carlo algorithm is more efficient as it calls the reward function only once:

  ```
  mc_shap_batch = MonteCarloShapleyBatch(x=x, fc=fc, ref=reference
          , n_iter=1000)
  ```

- Projected Stochastic Gradient Descent algorithm:

  ```
  sgd_est = SGDshapley(d, C=y.max())
  sgd_shap = sgd_est.sgd(x=x, fc=fc, ref=reference
          , n_iter=5000, step=.1, step_type="sqrt")
  ```

[1] `https://libraries.io/pypi/shapkit`

4 Application: explanation of a machine learning model that performs detection network intrusion

We are interested by the detection network intrusions protects a computer network from unauthorized users, including perhaps insiders. The objective is to build a predictive model capable of distinguishing between illegitimate (intrusions) and legitimate connections and to be able to understand the prediction made for an observation. This second task is important for the operational who uses the model, to be able to detect some false positive and effectively categorise the attacks. We use the KDD Cup 1999 Data [2], which includes a wide variety of intrusions simulated in a military network environment. The features are the basic features of individual TCP connections (e.g. number of seconds of the connection, type of protocol, etc.), some content features within a connection suggested by domain knowledge (e.g. number of failed login attempts, number of "root" accesses, etc.) and traffic features computed using a two-second time window (e.g. number of connections to the same host as the current connection in the past two seconds, number of connections to the same service as the current connection in the past two seconds, etc.). We focus on DoS attacks, without distinguishing between the different categories of DoS attacks. Denial-of-service attack is a cyber-attack in which the attacker seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled. The data are labelled: one class for the normal connection and the second for DoS attack, whatever the type of attacks it is (smurf, syn flood, etc.). A ML model is trained to distinguish between normal connection and DoS attacks. In this talk, we will use Shapley Values with two objectives: understand what are the important elements that leads to an alert and use these elements to try to refine the characterization of the attack undergone.

Acknowledgement

This work is supported by the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

[2] `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`

References

[1] Molnar. *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*. 2019.

[2] Barredo Arrieta, Diaz Rodriguez, Del Ser, Bennetot, Tabik, Barbado González, Garcia, Gil-Lopez, Molina, Benjamins, Chatila, and Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 2019.

[3] Ribeiro, Singh, and Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[4] Shapley. A value for n-person games. In *Contributions to the Theory of Games*. 1953.

[5] Lundberg and Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] Štrumbelj and Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, December 2014.

[7] Strumbelj and Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, 11:1–18, March 2010.

[8] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. *arXiv:1306.4265 [cs]*, February 2014. arXiv: 1306.4265.

[9] Grah and Thouvenot. A projected sgd algorithm for estimating shapley value applied in attribute importance. 2020.

[10] Castro, Gómez, Molina, and Tejada. Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. *Computers and Operations Research*, 2017.

[11] Aas, Jullum, and Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv:1903.10464 [cs, stat]*, February 2020. arXiv: 1903.10464.

Correcting bias sampling using weigthed empirical risk minimisation in statistical learning

Charles Tillier [1] & Robin Vogel [2] & Mastane Achab [3] & Stéphan Clémençon [4]

[1] *University of Versailles, 45 avenue des états-unis, 78000 Versailles, France*
[1] *charles.tillier@gmail.com*
[2 3 4] *Télécom-Paris, 19 Place Marguerite Perey, 91120 Palaiseau*
[2 3 4] *first.last@telecom-paris.fr*

Résumé. On considère un problème de minimisation du risque en apprentissage statistique où la distribution P' de l'échantillon d'apprentissage diffère de celle du test P. Lorsque le ratio de vraisemblance $\Phi = dP/dP'$ est connu, on montre que la procédure du minimisation du risque empirique peut être étendue à ce problème de *transfer learning* en utilisant les mêmes idées que celles de l'échantillonnage préférentiel. Lorsque Φ est inconnu, en pratique, la vraisemblance a souvent une forme simple explicite et peut être estimée directement à l'aide des Z_i' et d'une information auxiliaire sur la population P. On montre alors que la capacité de généralisation de notre approche est conservée en injectant les estimations $\Phi(Z_i')$ dans la version pondérée du risque empirique. On présente quelques applications numériques pour mettre en avant le potentiel de la méthode.

Mots-clés. Apprentissage statistique, Risque empirique, Echantillonnage préférentiel

Abstract. We consider statistical learning problems when the distribution P' of the training set differs from the test distribution P involved in the risk one seeks to minimize. When the likelihood ratio $\Phi(z) = dP/dP'(z)$ is known, we show that Empirical Risk Minimization (ERM) approach extends to this specific *transfer learning* setup using the same idea as that behind Importance Sampling. When Φ is unknown, we show that in practice it often takes a simple form and can be directly estimated from the Z_i''s and some auxiliary information on the statistical population P. We then prove that the generalization capacity of the approach aforementioned is preserved when plugging the resulting estimates of the $\Phi(Z_i')$'s into the weighted empirical risk. Numerical results provide empirical evidence of the relevance of our approach.

Keywords. Empirical risk minization, Importance sampling, Transfer learning

1 The framework of ERM

The standard empirical risk minimization problem (ERM) is the following: Z is a random variable (rv) valued in a space $\mathcal{Z}$ with distribution P, Θ is a parameter space and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is a loss function. Then, the risk one seeks to minimize is : $\forall \theta \in \Theta$,

$$\mathcal{R}_P(\theta) = \mathbb{E}_P\left[\ell(\theta, Z)\right]. \tag{1.1}$$

In practice P is unknown and learning is based on the sole observation of an independent and identically distributed (iid) sample $Z_1, \ldots, Z_n$ drawn from P and the risk (1.1) may be replaced by its empirical counterpart:

$$\widehat{\mathcal{R}}_P(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta), \tag{1.2}$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{Z_i}$ is the empirical measure of P and δ is the Dirac measure. The performance of minimizers of (1.2) can be studied by means of concentration inequalities, quantifying the fluctuations of the maximal deviations $\sup_{\theta\in\Theta} |\widehat{\mathcal{R}}_P(\theta) - \mathcal{R}_P(\theta)|$ under complexity assumptions for the functional class $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$, see [3]. The poor control of the data acquisition process in the big data era leads to the risk of jeopardizing the generalization ability of the algorithms; see [4]. In particular, bias selection issues in machine-learning are now the subject of much attention in the literature, see *e.g.* [1], [5], [6] and [7], since it creates discriminations in the algorithms, see [2]. To tackle this problem, we consider the case where the iid sample $Z'_1, \ldots, Z'_n$ available for training is not drawn from P but from another distribution P', with respect to which P is absolutely continuous, and the goal pursued is to set theoretical grounds for the application of ideas behind Importance Sampling (IS) methodology to extend the ERM approach to this learning setup.

2 Weighted ERM (WERM)

We start by investigating conditions guaranteeing that values for the parameter θ that nearly minimize (1.1) can be obtained through minimization of a weighted version of the empirical risk based on the Z'_i's, namely

$$\widetilde{\mathcal{R}}_{w,n}(\theta) = \mathcal{R}_{\widetilde{P}_{w,n}}(\theta), \tag{2.1}$$

where $\widetilde{P}_{w,n} = (1/n)\sum_{i=1}^n w_i \delta_{Z_i'}$ and $w = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ is a certain weight vector. Of course, ideal weights w^* are given by the likelihood function $\Phi(z) = (dP/dP')(z)$: $w_i^* = \Phi(Z_i')$ for $i \in \{1, \ldots, n\}$ and in this simple case $\mathbb{E}_{P'}[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)] = \mathcal{R}_P(\theta)$, so that generalization bounds for the $\mathcal{R}_P$-risk excess of minimizers of the empirical risk with ideal weights can be directly established by studying the concentration properties of the empirical process related to the Z_i''s and the class of functions $\{\Phi(\cdot)\ell(\theta, \cdot) : \theta \in \Theta\}$ (see section 2.1). However, the *importance function* Φ is unknown in general, just like distribution P. As highlighted in Section 2.2, in far from uncommon situations, the (ideal) weights w_i^* can be estimated from the Z_i's combined with auxiliary information on the target population P.

2.1 Known importance function

Consider the situation where Φ is known, insofar as we shall subsequently develop techniques aiming at mimicking the minimization of the ideally weighted empirical risk

$$\widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n}\sum_{i=1}^n w_i^* \ell(\theta, Z_i'), \tag{2.2}$$

namely the (unbiased) IS estimator of (1.1) based on the instrumental data $Z_1', \ldots, Z_n'$. The following result describes the performance of minimizers $\widetilde{\theta}_n^*$ of (2.2).

Lemma 2.1 *Assuming that ℓ and Φ are both bounded functions, with probability at least $1-\delta$, we have: $\forall n \geq 1$,*

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 4||\Phi||_\infty \mathbb{E}\left[R_n'(\mathcal{F})\right] + 2||\Phi||_\infty \sup_{(\theta,z)\in\Theta\times\mathcal{Z}} \ell(\theta, z)\sqrt{\frac{2\log(1/\delta)}{n}}$$

where $R_n'(\mathcal{F})$ denotes the Rademacher average associated to the class of function $\mathcal{F}$.

Note that when $P = P'$, we have $\Phi \equiv 1$ and the bound stated in Lemma 2.1 simply describes the performance of standard empirical risk minimizers.

2.2 Unkown importance function

While it is clear that the situation where the likelihood ratio is known is unrealistic, as illustrated by the example below, in many statistical learning problems with biased

training distribution, Φ takes a simplistic form and can be easily estimated from the Z_i''s combined with auxiliary information on P.

Binary classification with varying class probabilities. Let $Z = (X, Y)$, Y being a binary variable valued in $\{-1, +1\}$ and the rv X takes its values in a measurable space $\mathcal{X}$ and models some information hopefully useful to predict Y. The parameter space Θ is a set $\mathcal{G}$ of measurable mappings $g : \mathcal{X} \to \{-1,\ +1\}$ and the loss function is given by $\ell(g,\ (x,y)) = \mathbb{I}\{g(x) \neq y\}$ for all g in $\mathcal{G}$ and any $(x,y) \in \mathcal{X} \times \{-1,\ +1\}$. The distribution P of the random pair (X,Y) may be described by the triplet (p, F_+, F_-) where $p = \mathbb{P}\{Y = +1\}$ and $F_\sigma(dx)$ is X's conditional distribution given $Y = \sigma 1$ with $\sigma \in \{-,\ +\}$. We assume the common hypothesis $p' < p$ where p is supposed to be known. The likelihood function takes the simple following form $\Phi(x,y) = \mathbb{I}\{y = +1\}\frac{p}{p'} + \mathbb{I}\{y = -1\}\frac{1-p}{1-p'} \stackrel{def}{=} \phi(y)$, which reveals that it depends on the label y solely, and the ideally weighted empirical risk process is

$$\widetilde{\mathcal{R}}_{w^*,n}(g) = \frac{p}{p'}\frac{1}{n}\sum_{i:Y_i'=1} \mathbb{I}\{g(X_i') = -1\} + \frac{1-p}{1-p'}\frac{1}{n}\sum_{i:Y_i'=-1} \mathbb{I}\{g(X_i') = +1\}. \tag{2.3}$$

In general the theoretical rate p' is unknown and one replaces (2.3) with

$$\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) = \frac{p}{n'_+}\sum_{i:Y_i'=1} \mathbb{I}\{g(X_i') = -1\} + \frac{1-p}{n'_-}\sum_{i:Y_i'=-1} \mathbb{I}\{g(X_i') = +1\}, \tag{2.4}$$

where $n'_+ = \sum_{i=1}^n \mathbb{I}\{Y_i' = +1\} = n - n'_-$, $\widehat{w}_i^* = \widehat{\phi}(Y_i')$ and $\widehat{\phi}(y) = \mathbb{I}\{y = +1\}np/n'_+ + \mathbb{I}\{y = -1\}n(1-p)/n'_-$. The stochastic process above is not a standard empirical process but a collection of sums of two ratios of basic averages. However, the following result provides a uniform control of the deviations between the ideally weighted empirical risk and that obtained by plugging the empirical weights into the latter.

Lemma 2.2 *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $p' \in (\varepsilon,\ 1-\varepsilon)$. For any $\delta \in (0,1)$, we have with probability larger than $1-\delta$:*

$$\sup_{g\in\mathcal{G}} \left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g)\right| \leq \frac{2}{\varepsilon^2}\sqrt{\frac{\log(2/\delta)}{2n}},$$

as soon as $n \geq 2\log(2/\delta)/\varepsilon^2$.

Consequently, minimizing (2.4) nearly boils down to minimizing (2.3). Combining Lemmas 2.2 and 2.1, we immediately get the generalization bound stated in the result below.

Theorem 2.3 *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $p' \in (\varepsilon,\ 1-\varepsilon)$. Let $\widetilde{g}_n$ be any minimizer of $\widetilde{\mathcal{R}}_{\widehat{w}^*,n}$ over class $\mathcal{G}$. For any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g\in\mathcal{G}} \mathcal{R}_P(g) \leq \frac{2\max(p,1-p)}{\varepsilon}\left(2\mathbb{E}[R'_n(\mathcal{G})] + \sqrt{\frac{2\log(2/\delta)}{n}}\right) + \frac{4}{\varepsilon^2}\sqrt{\frac{\log(4/\delta)}{2n}},$$

as soon as $n \geq 2\log(4/\delta)/\varepsilon^2$; where $R'_n(\mathcal{G}) = (1/n)\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}} |\sum_{i=1}^n \sigma_i \mathbb{I}\{g(X'_i) \neq Y'_i\}|]$.

Such theoritical control of the weighted version of the ERM extends to numerous pratical cases such as multiclass classification, Positive-Unlabeled learning, et cetera.

3 Numerical experiments

A natural extension of binary classification is multiclass classification in a stratified population : Y and Y' take values in $\{1,\ \ldots,\ J\}$, $J \geq 1$, and each labeled observation (X,Y) belongs to a certain random stratum S in $\{1,\ \ldots,\ K\}$ with $K \geq 1$. Again, the distribution P of a random element $Z = (X,Y,S)$ may be described by the parameters $\{(p_{j,k}, F_{j,k}) :\ 1 \leq j \leq J,\ 1 \leq k \leq K\}$ where $F_{j,k}$ is the conditional distribution of X given $(Y,S) = (j,k)$ and $p_{j,k} = \mathbb{P}_{(X,Y,S)\sim P}\{Y = j, S = k\}$. Then, we have

$$dP(x,y,s) = \sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{I}\{y = j, s = k\} p_{j,k} dF_{j,k}(x),$$

and considering a distribution P' with $F_{j,k} \equiv F'_{j,k}$ but possibly different class-stratum probabilities $p'_{j,k}$, the likelihood function becomes

$$\frac{dP}{dP'}(x,y,s) = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{p_{j,k}}{p'_{j,k}} \mathbb{I}\{y = j, s = k\} \stackrel{def}{=} \phi(y,s).$$

Note that Theorem 2.3 extends to the latter case. We now focus on the MNIST dataset consisting of 60000 images for training and 10000 images for testing, labels being the value of the digits so that $K = 10$. There is no class bias in the original dataset so bias between classes is induced (see Figure (1)) using the power law strategy as follows

$$p'_k = \frac{\gamma^{-\frac{\lfloor K/2 \rfloor}{\sigma(k)}} p_k}{\sum_{l=1}^{K} \gamma^{-\frac{\lfloor K/2 \rfloor}{\sigma(k)}} p_k}$$

where $0 \leq \gamma \leq 1$ is a strata bias parameter. The optimization dynamics are summarized in Figure (2). We report the median over 100 runs of these values for the test set and a fixed random sample of the train set. For the test set, we represent 95% confidence-intervals in a lighter tone. The x-axis corresponds to the number of iterations of the learning process. It appears that for the uniform weights, we see that the misclassification rate is pretty low for the train set, but poor for the test set. By reweighting the instances, we favor low error over the test set, which gives a miss probability reduced by half.

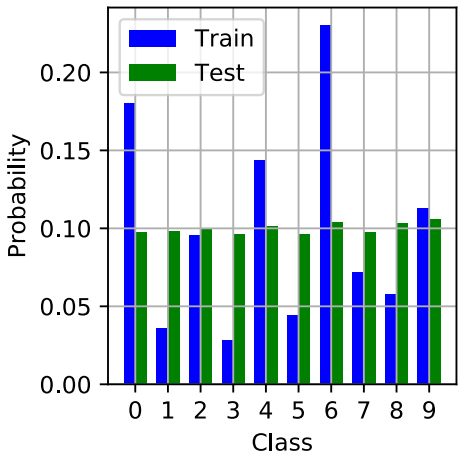


Figure 1: Comparison p_k's and p'_k's.

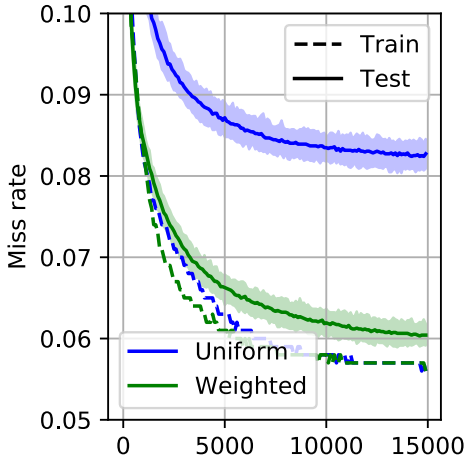


Figure 2: Dynamics for the class reweighting experiment with MNIST.

Bibliographie

[1] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V. and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NIPS*, 29, 4349-4357.

[2] Zafar, M. B., Valera, I., Gomez-Rodriguez, M. and Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *JMLR*, 20(75), 1-42.

[3] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. ESAIM: probability and statistics, 9, 323-375.

[4] Devroye, L., Gyorfi, L. and Lugosi, G. (2013). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science and Business Media.

[5] Zhang, K., Schölkopf, B., Muandet, K. and Wang, Z. (2013). Domain adaptation under target and conditional shift. *International Conference on Machine Learning* (pp. 819-827).
[6] Liu, Z., Yang, J.-a., Liu, H., and Wang, W. (2016). Transfer learning by sample selection bias correction and its application in communication specific emitter identification. *Journal of Communication*, 11, 417-427.
[7] Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 601-608.

Choix de la loi d'intensité dans les mélanges de Poisson basé sur la théorie des valeurs extrêmes

Samuel Valiquette[1,2,3,4] & Frédéric Mortier[1] & Jean Peyhardi[2] & Gwladys Toulemonde[2,3]

[1] *CIRAD, UPR Forêts et Sociétés, F-34398 Montpellier, France ; frederic.mortier@cirad.fr.*

[2] *IMAG, CNRS, Université de Montpellier ; gwladys.toulemonde@umontpellier.fr, jean.peyhardi@umontpellier.fr.*

[3] *LEMON, Inria*

[4] *Université de Sherbrooke ; samuel.valiquette@usherbrooke.ca.*

Résumé. Les mélanges de Poisson ont de nombreuses applications pour plusieurs champs de recherche. Par exemple, en écologie, ils permettent de modéliser des populations d'espèces en présence de surdispersion. Il existe cependant une infinité de mélanges possibles. Ce travail propose une stratégie pour choisir la loi de mélange à l'aide de la théorie des valeurs extrêmes. Pour ce faire, on présente les résultats associés à la théorie des valeurs extrêmes dans le cas discret ainsi que les catégories de mélanges possibles. Finalement, on propose à l'aide de ces résultats un arbre de décision permettant de sélectionner quel type de mélange de Poisson est adéquat pour l'ajustement des données.

Mots-clés. Mélange Poisson, Théorie des valeurs extrêmes discrètes, Méthode des excès.

Abstract. Poisson mixtures are widely used in many fields of research. For example, in ecology, they are used to model overdispersed species' population. However, there is a lot of possibilities to choose from. We present in this paper a strategy to adequately choose the mixed distribution using extreme value theory. We present this theory in the discrete case and apply it to Poisson mixtures. Finally, we suggest a decision tree that will allow to select what type of mixture is adequate for the data.

Keywords. Poisson mixture, Discrete extreme value theory, Peak-over-threshold.

1 Introduction

En écologie, modéliser et prédire la distribution des espèces est essentiel pour anticiper l'impact des changements climatiques et des pressions humaines sur la diversité des écosystèmes. Une approche est de supposer que l'abondance d'une espèce est distribuée selon une loi de Poisson dont l'intensité dépend de caractéristiques environnementales. Cette dernière requiert cependant une forte hypothèse : l'égalité entre l'espérance et la variance. Or, en raison de différents facteurs (dispersion limitée, compétition entre espèces ou autres), les données d'abondance présentent régulièrement une surdispersion ce qui tend à violer cette propriété. Cette surdispersion se manifeste soit par un excès de zéros,

soit par des valeurs extrêmes, les deux phénomènes n'étant pas exclusifs. Dans ce cas, un modèle de Poisson s'ajustera mal aux données.

Une stratégie, pour prendre en compte cette surdispersion, repose sur l'utilisation de modèles de mélanges finis ou non (Karlis et Xekalaki, 2005). Celle-ci consiste à supposer que l'intensité de la loi de Poisson, λ, n'est plus une valeur fixe (inconnue) mais est, elle-même, aléatoire. De tels modèles permettent d'avoir une plus grande variance et sont en mesure d'observer plus de zéros ou des valeurs extrêmes. Karlis et Xekalaki (2005) listent jusqu'à 30 exemples de modèles de Poisson en mélange selon le choix de la distribution du paramètre λ. D'un point de vue général, toutes lois dont le support est positif sont candidates. À l'heure actuelle, il ne semble pas exister d'études et de travaux permettant de choisir la ou les distributions les mieux adaptées. Ainsi, ce travail a pour but de répondre à ce besoin en proposant une méthode permettant de choisir, au regard des données, une famille de lois adéquates. Précisément, nous proposons dans ce travail une approche permettant de choisir les lois de mélange en se référant au comportement en queue de la distribution des observations. Pour cela, on rappelle des éléments de la théorie des valeurs extrêmes et de la situation particulière du cadre discret, l'application de cette théorie aux mélanges de Poisson et enfin nous présentons une stratégie de choix de mélange.

2 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes s'intéresse à l'étude de valeurs très grandes et rares et propose des méthodes paramétriques pour analyser le comportement en queue d'une distribution quelconque. Une approche classique est basée sur l'étude des excès d'un échantillon. Plus précisément, soit $X \sim F$ une variable aléatoire (continue ou discrète) dont le support est fini ou non. On appelle un excès, la quantité $X - u|X > u$ où u est un seuil fixé. Pickands (1975) démontre que s'il existe une suite $\sigma(u)$ strictement positive telle que $\sigma(u)^{-1}(X-u)|X > u$ converge vers une distribution non-dégénérée H lorsque u tend vers le point terminal du support, alors H sera une distribution Pareto généralisée (GPD). Cette distribution est définie par sa fonction de survie :

$$\overline{H}_{\gamma,\sigma}(y) = \begin{cases} \left(1+\gamma\frac{y}{\sigma}\right)^{-1/\gamma} & \text{si } \gamma \neq 0 \\ \exp\left(-\frac{y}{\sigma}\right) & \text{sinon} \end{cases}$$

dont le support est $\mathbb{R}^+$ si $\gamma \geq 0$ ou $\left[0; -\frac{\sigma}{\gamma}\right]$ si $\gamma < 0$, où σ et γ sont les paramètres d'échelle et de forme respectivement. Dans le cas où la convergence a lieu, on dit que la distribution de X est dans le domaine d'attraction des maximums. Il existe trois domaines d'attraction, selon que γ soit négatif, nul ou positif. Ces trois domaines se nomment respectivement le domaine de Weibull, de Gumbel et de Fréchet. Chaque domaine caractérise le comportement en queue de façon unique (Resnick, 1987). Dans ce travail, seuls les domaines de Gumbel et Fréchet, dont le support est infini, seront spécifiquement étudiés.

Bien que la théorie des valeurs extrêmes soit très générale, on peut constater des différences lorsqu'on se place dans le cas continu ou discret. Alors qu'il est relativement aisé de démontrer l'existence d'une suite normalisante dans le cas continu, cela n'est plus vrai dans le cas discret. Les contre exemples sont nombreux. On peut citer parmi d'autres les lois de Poisson, binomiale négative, ou géométrique. Anderson (1970) puis Shimura (2012) ont étudié ce phénomène et démontrent qu'une condition nécessaire afin qu'une loi discrète F soit dans un domaine d'attraction est que celle-ci soit à queue longue. Cette propriété est définie par la limite $\frac{1-F(x+1)}{1-F(x)} \to 1$ lorsque $x \to \infty$. Or cette condition n'est pas satisfaite en général. Cependant, lorsque cette limite existe et tend vers une valeur entre $(0, 1)$, Anderson (1970) et Shimura (2012) ont démontré que de telles lois seront "proches" du domaine de Gumbel. Autrement dit, ils démontrent que si la variable aléatoire discrète était, en fait, continue, alors on se retrouverait dans le domaine Gumbel. La stratégie qu'on propose repose sur ce résultat clé.

3 Mélanges Poisson et domaine d'attraction

Pour choisir la loi de mélange, on souhaite connaître le comportement en queue du mélange résultant selon le domaine d'attraction de la loi de λ. Perline (1998) propose deux classes de mélanges associées aux domaine de Fréchet et Gumbel. Dans la première, il démontre que si on utilise une loi dans le domaine de Fréchet satisfaisant une condition suffisante nommée Von Mises (Resnick, 1987), alors le mélange sera également dans le domaine de Fréchet. Plusieurs lois classiquement utilisées dans le domaine de Fréchet satisfont cette condition, par exemple l'inverse-gamma ou la demi-Cauchy. Pour la deuxième classe, Perline (1998) démontre que si la loi de λ est dans Gumbel et satisfait aussi la condition suffisante ainsi qu'une condition sur le taux de défaillance, alors le mélange Poisson sera aussi dans Gumbel. À titre d'exemple, la log-normale permettra au mélange d'être dans cette classe. On nommera respectivement ces deux classes de mélanges par Fréchet et Gumbel.

Cependant, plusieurs lois dans le domaine de Gumbel ne produiront pas de mélange qui seront éléments de cette deuxième classe. Par exemple, la loi gamma satisfait seulement la condition de Von Mises et lorsqu'on l'utilise pour λ, le mélange Poisson produira une binomiale négative. Cette dernière, malheureusement, ne possède aucun domaine d'attraction. Or la binomiale négative est classiquement usitée pour l'analyse de données de la comptage et il serait important d'avoir une classe de mélange qui la décrit. Pour ce faire, on a étudié les mélanges qui utilisent des lois similaires à la gamma. On dit qu'une loi continue avec densité f a un comportement gamma si

$$\lim_{x\to\infty} \frac{f(x)}{C(x)x^{\alpha}e^{-\beta x}} = 1$$

où $C(x)$ est localement bornée sur $(0, \infty)$ et à variation lente, c'est-à-dire $\lim_{t\to\infty} \frac{C(tx)}{C(t)} = 1$

(Bingham et al., 1987), $\alpha \in \mathbb{R}$ et $\beta > 0$.

Cette définition donnée par Willmot (1990) inclut, entre autre, l'inverse-Gaussienne, l'exponentielle et, évidemment, la gamma. Avec ce type de lois, on peut décrire le comportement de plusieurs mélanges qui ne sont pas inclus dans les classes de Perline (1998). En effet, on démontre deux résultats clés : les lois à comportement gamma sont dans le domaine d'attraction de Gumbel et le mélange Poisson résultant n'aura aucun domaine d'attraction. Précisément on montre le théorème suivant :

Théorème 1 (Valiquette (2020))**.** *Soit X un mélange Poisson de distribution F_X, supposons que la loi de λ, notée F, a un comportement gamma, alors :*

1. *F est dans le domaine de Gumbel.*
2. *Pour tout entier k, $\lim_{x\to\infty} \frac{1-F_X(x+k)}{1-F_X(x)} = \left(\frac{1}{1+\beta}\right)^k \in (0,1)$. En particulier, pour $k = 1$, F_X n'est pas à queue longue et donc ne possède aucun domaine d'attraction.*

Malgré l'absence de domaine d'attraction, ces mélanges peuvent être décrits par leurs comportement en queue. En effet, comme mentionné en section 2, ces mélanges seront, dans un sens, "proches" du domaine de Gumbel. Ainsi, on propose d'ajouter une troisième classe à celles de Perline (1998) qu'on nomme Pseudo-Gumbel. Ces trois classes permettront de choisir le mélange adéquat aux données de comptage. On donne quelques exemples de mélange et leur classe selon la loi de λ en table 1.

Loi sur λ	Domaine (Loi)	Mélange	Classe (Mélange)
Demi-Cauchy	Fréchet	Poisson demi-Cauchy	Fréchet
Inverse-gamma	Fréchet	Poisson inverse-gamma	Fréchet
Log-Gaussienne	Gumbel	PLG	Gumbel
Weibull(α, β)	Gumbel	Poisson-Weibull	Gumbel (si $\alpha < \frac{1}{2}$)
Gamma(α,β)	Gumbel	Binomiale négative	Pseudo-Gumbel
Inverse-Gaussienne	Gumbel	Sichel	Pseudo-Gumbel

TABLE 1 – Domaines d'attraction. Quelques exemples

4 Choix de mélange Poisson

L'idée principale de notre stratégie est d'utiliser les trois classes de mélanges afin de bien choisir la loi pour le paramètre d'intensité λ. Dans cette section, on décrit en détail les étapes à suivre et on peut visualiser cette stratégie à l'aide d'un arbre de décision (voir Figure 1). Pour cela, on commence par identifier si les données de comptage appartiennent à un domaine d'attraction. Pour vérifier cette propriété, il faut fixer un seuil suffisamment élevé et ajuster une GPD aux excès. On aura deux situations possibles : l'ajustement

des excès est adéquat ou non. Il existe plusieurs méthodes pour procéder à une telle vérification. Par exemple cela peut être un graphique Quantile-Quantile (Coles, 2001) ou un test d'ajustement bootstrap (Villaseñor-Alva et Gonzàlez-Estrada, 2009). Dans la première situation, cela signifie que les données possèdent un domaine d'attraction. Dans ce cas, il serait préférable d'utiliser une loi pour λ qui permette de préserver le domaine d'attraction. Ceci revient à utiliser des lois qui correspondent aux classes de Perline (1998). Si l'estimation est adéquate avec $\gamma = 0$, on devrait choisir une loi qui permet de rester dans Gumbel tel que la log-normale. Dans le cas où l'estimation est strictement positive, alors il faudra choisir une loi dans le domaine de Fréchet comme la demi-Cauchy.

Cependant si on se retrouve dans la deuxième situation, c'est-à-dire les excès ne sont pas ajustés adéquatement par la GPD, alors les données discrètes ne possèdent pas de domaine d'attraction. Or, il serait possible qu'elles soient "proches" du domaine de Gumbel. Pour tester si c'est bien le cas, on ajoute aux données un bruit aléatoire d'une uniforme(0,1) afin de les rendre continues. Cette technique se nomme *jittering* en anglais et possède de nombreuses applications (Nagler, 2018). En réajustant une GPD à ces nouveaux excès continus avec un paramètre de forme fixé à $\gamma = 0$, on vérifie si cette fois c'est adéquat. Si c'est bien le cas, on se retrouve dans la classe Pseudo-Gumbel et l'utilisation de loi à comportement gamma est recommandée. Par exemple l'utilisation de l'inverse-Gaussienne pour λ permet d'être dans cette situation. Si au contraire l'ajustement n'est toujours pas adéquat, alors il faudra procéder à une analyse différente pour choisir le mélange.

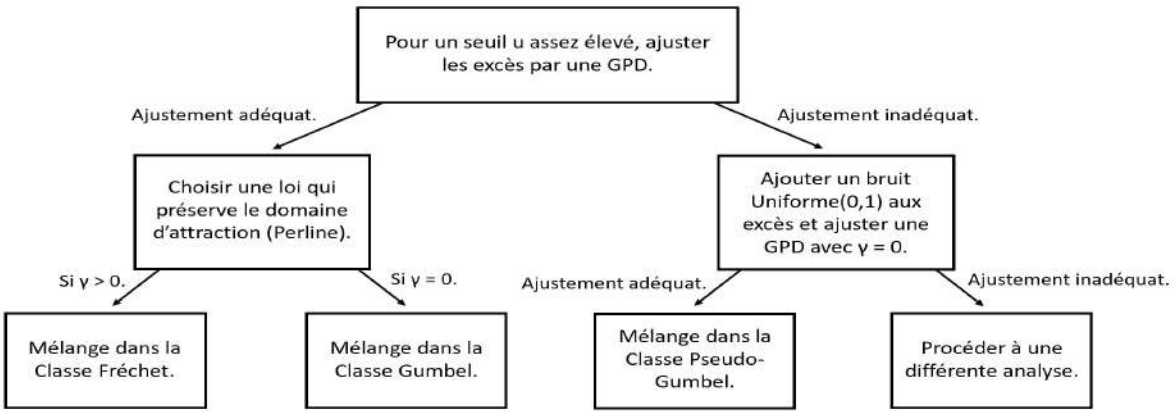


FIGURE 1 – Arbre de décision pour choisir les lois de mélange potentielles

5 Discussion et perspectives

L'arbre de décision a été validé par simulation en utilisant plusieurs lois de mélanges selon les différentes classes. Pour différents représentants des trois catégories, la pratique semble concorder avec la théorie lorsqu'on étudie des lois dont la variance est raisonnable. Néanmoins, lorsqu'on étudie par exemple la log-normale avec des paramètres d'échelle qui produisent une très forte variance, il n'est pas rare de conclure que cette loi particulière appartient au domaine de Fréchet et non à celui de Gumbel. Ainsi, il se pourrait en pratique que l'on confonde ces deux classes lorsque la variance est trop importante. Cet aspect devrait être étudié plus en détail, par exemple en ajustant directement sur un ensemble de données simulées et sur des cas d'étude. Finalement, il serait pertinent d'ajouter à cette approche des covariables et d'étudier si celle-ci peut être étendue au cadre de la régression.

Remerciements : Ce travail a été soutenu par le projet GAMBAS financé par l'agence nationale de la recherche (ANR-18-CE02-0025).

Bibliographie

Anderson, C.W. (1970). Extreme Value Theory for a Class of Discrete Distributions with Applications to some Stochastic Processes, *Journal of Applied Probability*, 7, pp. 99-113.
Bingham, N.H., Goldie C.M. and Teugles J.L. (1987). Regular Variation, *Cambridge University Press*.
Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values, *Springer*.
Karlis, D. and Xekalaki, E. (2005). Mixed Poisson Distributions, *International Statistical Review*, 73, pp. 35-58.
Nagler, T. (2018). A Generic Approach to Nonparametric Function Estimation with Mixed Data, *Statistics and Probability Letters*, 137, pp. 326-330.
Perline, R. (1998). Mixed Poisson Distributions Tail Equivalent to their Mixing Distributions, *Statistics and Probability Letters*, 38, pp. 229-233.
Shimura, T. (2012). Discretization of Distributions in the Maximum Domain of Attraction, *Extremes*, 15(3), pp. 299-317.
Valiquette, S. (2020). Théorie des Valeurs Extrêmes dans le Cadre des Mélanges de Poisson, URL https ://gambas.cirad.fr/products/reports
Villaseñor-Alva, J.A. and Gonzàlez-Estrada, E. (2009). A Bootstrap Goodness of Fit Test for the Generalized Pareto Distribution, *Computational Statistics and Data Analysis*, 53, pp. 3835-3841.
Willmot, G.E. (1990). Asymptotic Tail Behaviour of Poisson Mixture with Applications, *Advances in Applied Probability*, 22, pp. 147-159.

Comportement asymptotique de tests de Sobolev sur la sphère unité.

Christine Cutting, Davy Paindaveine and Thomas Verdebout

ECARES and Mathematics Department, Université Libre de Bruxelles, Boulevard du Triomphe, CP210, B-1050 Brussels, Belgium, email: tverdebo@ulb.ac.be.

Résumé. Dans ce travail, nous considérons le problème de test d'uniformité sur l'hypersphère unité de $\mathbb{R}^p$. Nous obtenons de nouveaux résultats sur le comportement asymptotique de tests de Sobolev sous des contre-hypothèses locales à symmétrie rotationnelle.

Mots-clés. Données directionnelles, tests d'uniformité, tests de Sobolev.

Abstract. In this work, we tackle the problem of testing uniformity on the unit hypersphere of $\mathbb{R}^p$. We obtain new results on the asymptotic behavior of Sobolev tests under rotationally symmetric alternatives.

Keywords. Directional data, uniformity tests, Sobolev tests.

1 Directional data and testing for uniformity

Directional statistics are dealing with observations that belong to the unit hypersphere $\mathbb{S}^{p-1} := \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|^2 = \mathbf{u}'\mathbf{u} = 1\}$ of $\mathbb{R}^p$ or more generally on compact Riemannian manifolds. Instances of directional data happen in meteorology (wind directions), astronomy (directions of cosmic rays, positions of stars), paleomagnetism (remanence directions), biology (protein structure, studies of animal navigation), forest sciences (directions of wildfire propagation), medicine (head normal vectors), and text mining (quantitative representation of documents in high-dimensional hyperspheres), to cite but some. Classical monographs on directional statistics are Watson (1983) and Mardia and Jupp (2000); a recent book that overviews the usage of some modern methods in directional statistics is Ley and Verdebout (2017).

When modeling directional data, that is, unit-norm multivariate vectors, a first natural question is to ask whether the directions at hand are uniformly distributed or, on the contrary, whether there exist modes of variation significantly different from uniformity. On the basis of n i.i.d. observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ with common distribution P on $\mathbb{S}^{p-1}$, the problem we tackle in this work is the problem of testing $\mathcal{H}_0 : \mathrm{P} \equiv \mathrm{Unif}(\mathbb{S}^{p-1})$ against $\mathcal{H}_1 : \mathrm{P} \neq \mathrm{Unif}(\mathbb{S}^{p-1})$, where $\mathrm{Unif}(\mathbb{S}^{p-1})$ stand for the uniform probability measure on $\mathbb{S}^{p-1}$. We study in this work tests belonging to the class of Sobolev tests for this problem. Sobolev tests are introduced in the next Section.

2 Sobolev tests

The class of so-called *Sobolev tests* has been introduced by Beran (1968, 1969) and Gine (1975). Sobolev tests are obtained using the eigenfunctions of the *Laplace–Beltrami operator* (or *Laplacian*) Δ acting on $\mathbb{S}^{p-1}$. Using the n-tuple of observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$, a Sobolev test rejects the null hypothesis of uniformity $\mathcal{H}_0$ for large values of

$$S_n := \frac{1}{n} \sum_{i,j=1}^{n} \sum_{k=1}^{\infty} v_k^2 \langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle, \tag{2.1}$$

where $\mathbf{u} \to \mathrm{t}_k(\mathbf{u})$ is a mapping from $\mathbb{S}^{p-1}$ to the space of eigenfunctions associated with the kth non-zero eigenvalue of the Laplacian, the v_k's are weights and $\langle f, g \rangle := \int_{\mathbb{S}^{p-1}} f(\mathbf{u}) g(\mathbf{u}) \, \mathrm{d}\mu(\mathbf{u})$ denotes the inner product on $L^2(\mathbb{S}^{p-1}, \mu)$ (μ is the surface area measure on $\mathcal{S}^{p-1}$). An explicit form for $\langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle$ on $\mathbb{S}^{p-1}$ exists. More precisely, given $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{p-1}$,

$$\langle \mathrm{t}_k(\mathbf{u}), \mathrm{t}_k(\mathbf{v}) \rangle = \begin{cases} 2\cos(k \angle(\mathbf{u}, \mathbf{v})), & \text{if } p = 2, \\ \left(1 + \frac{2k}{p-2}\right) C_k^{(p-2)/2}(\mathbf{u}'\mathbf{v}), & \text{if } p > 2, \end{cases} \tag{2.2}$$

where $\cos \angle(\mathbf{u}, \mathbf{v}) = \mathbf{u}'\mathbf{v}$ and C_k^{α} denote the Gegenbauer polynomial of index α and order k. Well-known Sobolev tests are

- the *Rayleigh test*. Taking $v_1 = 1$ and $v_k = 0$ for $k \geq 2$ in (2.1) we obtain the Rayleigh test statistic on $\mathbb{S}^{p-1}$ given by

$$R_n = \frac{p}{n} \sum_{i,j=1}^{n} \mathbf{U}_i' \mathbf{U}_j. \tag{2.3}$$

  Under $\mathcal{H}_0$, R_n is asymptotically χ_p^2 distributed.

- the *Bingham test*. When $\mathbf{U} \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, then $\mathbb{E}[\mathbf{U}\mathbf{U}'] = \frac{1}{p}\mathbf{I}_p$. The Bingham test evaluates this latter sphericity property of $\mathbf{U}$ by the test statistic

$$B_n := \frac{np(p+2)}{2} \left(\mathrm{tr}(\mathbf{S}^2) - \frac{1}{p} \right),$$

  where $\mathbf{S} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{U}_i \mathbf{U}_i'$ is the empirical covariance matrix of the $\mathbf{U}_i$'s. Under $\mathcal{H}_0$, B_n is asymptotically $\chi^2_{(p-1)(p+2)/2}$ distributed. The statistic B_n is obtained by letting $v_2 = 1$ and $v_k = 0$ for $k \neq 2$ in (2.1).

While much is known about the asymptotic behavior of several Sobolev tests under the null hypothesis of uniformity and when the dimension is fixed, less is known about the asymptotic behaviour of such tests under local alternatives, even under the rotationally symmetric alternatives defined in the next section.

3 Asymptotic results for the Rayleigh test

In this section, we present several results obtained in Cutting *et al.* (2017). We consider specific alternatives to the null of uniformity over the p-dimensional unit sphere $\mathcal{S}^{p-1}$, namely rotationally symmetric alternatives. A p-dimensional unit vector $\mathbf{U}$ is said to be *rotationally symmetric about* $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ if and only if $\mathbf{OU}$ is equal in distribution to $\mathbf{U}$ for any orthogonal $p \times p$ matrix $\mathbf{O}$ satisfying $\mathbf{O}\boldsymbol{\theta} = \boldsymbol{\theta}$. We actually restrict to rotationally symmetric densities of the form

$$\mathbf{u} \mapsto c_{p,\kappa,f} f(\kappa\, \mathbf{u}'\boldsymbol{\theta}), \qquad \mathbf{x} \in \mathcal{S}^{p-1}, \tag{3.4}$$

where $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ is a location parameter, $\kappa(> 0)$ is a concentration parameter, and the function $f : \mathbb{R} \to \mathbb{R}^+$ is monotone strictly increasing, twice differentiable at 0, and satisfies $f(0) = f'(0) = 1$. Consider triangular arrays of observations $\mathbf{U}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ where the random vectors $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ take values in $\mathcal{S}^{p_n-1}$. More specifically, for any $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$, $\kappa_n > 0$ and f as above, we will denote as $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ the hypothesis under which $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ are mutually independent and share the common density $\mathbf{u} \mapsto c_{p_n,\kappa_n,f} f(\kappa_n\, \mathbf{u}'\boldsymbol{\theta}_n)$; $\mathrm{P}^{(n)}_0$ will denote triangular arrays of uniformly distributed observations. We have the following result.

Proposition 3.1 *Let (p_n) be a sequence in $\{2, 3, \ldots\}$. Let $(\boldsymbol{\theta}_n)$ be a sequence such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for all n, (κ_n) be a positive sequence such that $\kappa_n = O(\sqrt{\frac{p_n}{n}})$. Then, the sequence of alternative hypotheses $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ and the null sequence $\mathrm{P}^{(n)}_0$ are mutually contiguous.*

The sequences κ_n of the form $\kappa_n = O(\sqrt{\frac{p_n}{n}})$ therefore characterize the high-dimensional contiguous alternatives to the uniform distribution. We also obtain in Cutting *et al.* (2017) the following result.

Proposition 3.2 *Let (p_n) be a sequence in $\{2, 3, \ldots\}$ and let $(\boldsymbol{\theta}_n)$ be a sequence such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for all n. Let $\kappa_n = \tau_n \sqrt{p_n/n}$, where the positive sequence (τ_n) is $O(1)$ but not $o(1)$. Then, as $n \to \infty$ under $\mathrm{P}^{(n)}_0$,*

$$\log \frac{d\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}}{d\mathrm{P}^{(n)}_0} = \tau_n \Delta^{(n)}_{\boldsymbol{\theta}_n} - \frac{\tau_n^2}{2} + o_{\mathrm{P}}(1), \tag{3.5}$$

where $\Delta^{(n)}_{\boldsymbol{\theta}_n} := \frac{\sqrt{p_n}}{\sqrt{n}} \sum_{i=1}^n \mathbf{U}'_{ni}\boldsymbol{\theta}_n$ is asymptotically standard normal. In other words, the model $\{\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa,f} : \kappa \geq 0\}$ is locally asymptotically normal at $\kappa = 0$ with central sequence $\Delta^{(n)}_{\boldsymbol{\theta}_n}$, Fisher information 1, and contiguity rate $\sqrt{p_n/n}$.

Proposition 3.2 entails that the test $\phi^{(n)}_{\boldsymbol{\theta}_n}$ rejecting the null at asymptotic level α whenever

$$\Delta^{(n)}_{\boldsymbol{\theta}_n} = \sqrt{np_n}\, \bar{\mathbf{X}}'_n \boldsymbol{\theta}_n > z_\alpha \tag{3.6}$$

is locally asymptotically most powerful for the considered problem. The Rayleigh test is not locally and asymptotically optimal for testing uniformity against specified-$\boldsymbol{\theta}_n$ rotationally symmetric alternatives. It is actually blind to the contiguous alternatives. It detects alternatives with $\kappa_n = O(\frac{p_n^{3/4}}{\sqrt{n}})$. We show in Cutting *et al.* (2017) that it is locally and asymptotically most powerful for the unspecified-$\boldsymbol{\theta}_n$ problem within the class of rotation-invariant tests.

4 Perspectives

High-dimensional results dealing with the asymptotic behavior of the Bingham test have been obtained in Cutting *et al.* (2020). Our objective in a near future is to extend the results obtained in Cutting *et al.* (2017) and Cutting *et al.* (2020) to the entire class of Sobolev tests.

Bibliographie

Beran, R. J. (1968). Testing for uniformity on a compact homogeneous space. *Journal of Applied Probability*, 5, pp. 177-195.

Beran, R. J. (1969). Asymptotic theory of a class of tests for uniformity of a circular distribution. *Annals of Mathematical Statistics*, 40, pp. 1196-1206.

Cutting, C., Paindaveine, D., and Verdebout, T. (2017). Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives. *Annals of Statistics*, 45(3), pp. 1024-1058.

Cutting, C., Paindaveine, D., and Verdebout, T. (2020). Testing uniformity on high-dimensional spheres: the non-null behaviour of the Bingham test. *Submitted*.

Ley, C. and Verdebout, T. (2017). Modern Directional Statistics. Chapman and Hall/CRC.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Watson, G. S. (1983). *Statistics on Spheres*, volume 6 of University of Arkansas Lecture Notes in the Mathematical Sciences. John Wiley & Sons, New York.

Generalized Weibull-tail distributions

Mariia Vladimirova[1] & Julyan Arbel[1] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France*
{mariia.valdimirova, julyan.arbel, stephane.girard}@inria.fr

Résumé. Des variables aléatoires à queue de type Weibull sont définies comme des variables aléatoires positives dont la queue droite se comporte de manière similaire à celle d'une variable aléatoire de Weibull. Nous introduisons des variables aléatoires de queue Weibull généralisées, des extensions qui ont une propriété de stabilité. Nous étudions la préservation des paramètres de queue de ces variables aléatoires sous des opérations telles que la multiplication, la puissance et la combinaison linéaire.

Mots-clés. Lois à queue de type Weibull, Lois généralisé à queue de type Weibull, Sous-Weibull, Stabilité

Abstract. Weibull-tail random variables are defined as non-negative random variables whose right tail behaves similarly to that of a Weibull random variable. We introduce generalized Weibull-tail random variables, extensions which have a property of stability. We study the preservation of tail parameters under operations like multiplication, power, and linear combination.

Keywords. Weibull-tail, Generalized Weibull-tail, Sub-Weibull, Stability

1 Introduction

The study of the distributional tail behavior arises in many applied probability models of different areas, such as hydrology (Strupczewski et al., 2011), finance (Rachev, 2003) and insurance risk theory (Ahmad et al., 2020). Since in most cases exact distributions are not available, deriving asymptotic relationships for their tail probabilities becomes essential. In this context, an important role is played by so-called Weibull-tail distributions (Gardes et al., 2011; Gardes and Girard, 2016).

A random variable X is called *Weibull-tail* with tail parameter $\beta > 0$ if its cumulative distribution function F satisfies

$$\overline{F}(x) = 1 - F(x) = \mathrm{e}^{-x^{\beta} l(x)}, \quad \text{for } x > 0, \tag{1}$$

where $l(x)$ is a slowly-varying function, i.e. it is a positive function such that

$$\lim_{x \to \infty} \frac{l(tx)}{l(x)} = 1, \quad \text{for all } t > 0.$$

We note $X \sim \mathrm{WT}(\beta)$. The family of Weibull-tail distributions includes a variety of fundamental distributions such as Gaussian ($\beta = 2$), exponential and gamma ($\beta = 1$), Weibull ($\beta > 0$), to name a few.

Some of the commonly used techniques to study the tail behavior is to consider probability tail bounds such as sub-Gaussian, sub-Exponential, or their generalization to sub-Weibull distributions (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2018). A non-negative random variable is called sub-Weibull with tail parameter $\theta > 0$ if its survival function is upper-bounded by that of a Weibull distribution:

$$\overline{F}(x) \leq a\mathrm{e}^{-bx^{1/\theta}}, \quad \text{for } x > 0 \text{ and some } a, b > 0. \tag{2}$$

This property ensures the existence of the moment generating function as well as bounds on moments. In contrast, the Weibull-tail properties characterize the survival or density functions without a hand on moments.

While tail parameters in Equation (1) and (2) of Weibull-tail and sub-Weibull properties are different, we can find some connections. Notice that for any constants $a, b > 0$ there exists a slowly-varying function $l(x) = b - \frac{\log a}{x^\beta}$ so that $a\mathrm{e}^{-bx^\beta} = \mathrm{e}^{-x^\beta l(x)}$. It means that if random variable X is sub-Weibull with parameter $\theta = 1/\beta > 0$, satisfying Equation (2), then survival function of X is upper-bounded by a Weibull-tail distribution with tail parameter β, satisfying Equation (1). If random variable X is Weibull-tail with tail parameter β, then from Potter's bounds, for $a_1, a_2 > 0$ we have

$$a_1\mathrm{e}^{-x^{\beta_1}} \leq \overline{F}(x) = \mathrm{e}^{-x^\beta l(x)} \leq a_2\mathrm{e}^{-x^{\beta_2}},$$

or $\mathrm{WT}(\beta) \subset \mathrm{SubW}(1/\beta_2)$ and $\mathrm{WT}(\beta) \not\subset \mathrm{SubW}(1/\beta_1)$ for x big enough and $\forall(\beta_1, \beta_2)$ such that $1/\beta_2 < 1/\beta < 1/\beta_1$.

In this work, we study the properties of Weibull-tail random variables and introduce their stable extensions, i.e. generalized Weibull-tail random variables. Firstly, we show that multiplication by a constant doesn't change a random variable tail parameter (Lemma 2.1). A power of a Weibull-tail and generalized Weibull-tail random variable results into a distribution with a tail parameter divided by the power (Lemma 2.1). Further, Theorem 2.1 confirms that a sum of generalized Weibull-tail random variables (including the dependent ones) remains generalized Weibull-tail of tail parameter equal to the minimum among those of the terms. In addition, we consider a product of independent generalized Weibull-tail random variables in Theorem 2.2.

2 Weibull-tail properties

We begin by introducing the concept of a generalized random variable of the Weibull tail, which has an additional stability property:

Definition 2.1 (Generalized Weibull-tail)**.** A random variable X is called *generalized Weibull-tail* with tail parameter $\beta > 0$ if its survival function $\overline{F}$ is bounded by Weibull-tail functions of tail parameter β with possibly different slowly-varying functions l_1 and l_2:

$$\mathrm{e}^{-x^\beta l_1(x)} \leq \overline{F}(x) \leq \mathrm{e}^{-x^\beta l_2(x)}, \text{ for } x > 0. \tag{3}$$

We note $X \sim \mathrm{GWT}(\beta)$.

If $X \sim \mathrm{WT}(\beta)$ with slowly-varying function l, then by taking $l_1 = l_2 = l$, we have equality in Equation (3) and $X \sim \mathrm{GWT}(\beta)$. Therefore, Weibull-tail family of distributions is a subset of generalized Weibull-tail distributions of the same tail parameter: $\mathrm{WT}(\beta) \subset \mathrm{GWT}(\beta)$.

Lemma 2.1 (Power and multiplication by a constant)**.** *Let non-negative random variable X be Weibull-tail (generalized Weibull-tail) with tail parameter β, then aX^b is Weibull-tail (generalized Weibull-tail) with tail parameter $\frac{\beta}{b}$ for $a, b > 0$.*

Proof. For $a, b > 0$, the tail of $Y = aX^b$ is $\mathbb{P}\Big(aX^b \geq y\Big) = \mathbb{P}\Big(X \geq \big(\frac{y}{a}\big)^{1/b}\Big)$.

If X is Weibull-tail with tail parameter β, then $\mathbb{P}\left(X \geq x\right) = \mathrm{e}^{-x^\beta l(x)}$, where l is a slowly-varying function. It implies

$$\mathbb{P}\Big(aX^b \geq y\Big) = \mathrm{e}^{-y^{\beta/b}\tilde{l}(y)},$$

where $\tilde{l}(y) = \frac{l((y/a)^{1/b})}{a^{\beta/b}}$ is a slowly-varying function.

If X is generalized Weibull-tail with tail parameter β, then $\mathrm{e}^{-x^\beta l_1(x)} \leq \mathbb{P}\left(X \geq x\right) \leq \mathrm{e}^{-x^\beta l_2(x)}$, where l_1 and l_2 are slowly-varying functions. It implies

$$\mathrm{e}^{-y^{\beta/b}\tilde{l}_1(y)} \leq \mathbb{P}\Big(aX^b \geq y\Big) \leq \mathrm{e}^{-y^{\beta/b}\tilde{l}_2(y)},$$

where $\tilde{l}_i(y) = \frac{l_i((y/a)^{1/b})}{a^{\beta/b}}, i = 1, 2$ are slowly-varying functions. □

Theorem 2.1 (Sum of generalized Weibull-tail RVs)**.** *Let non-negative (possibly dependent) random variables X and Y be generalized Weibull-tail of parameters β_x and β_y. Then, $X + Y$ is generalized Weibull-tail of the parameter* $\min\{\beta_x, \beta_y\}$.

Proof. For any two distributions X and Y we have an upper bound

$$\mathbb{P}(X + Y \geq z) \leq \mathbb{P}(X \geq z/2 \cup Y \geq z/2) \leq \mathbb{P}(X \geq z/2) + \mathbb{P}(Y \geq z/2).$$

For non-negative random variables X and Y we have $\mathbb{P}(X + Y \geq z) \geq \mathbb{P}(X \geq z)$ and $\mathbb{P}(X + Y \geq z) \geq \mathbb{P}(Y \geq z)$. By combining these two inequalities we have a lower bound $\mathbb{P}(X + Y \geq z) \geq \max\{\mathbb{P}(X \geq z), \mathbb{P}(Y \geq z)\}$.

Thus, for survival function $\overline{F}_\Sigma$ of sum of random variables X and Y with survival functions $\overline{F}_X$ and $\overline{F}_Y$, the following inequality holds

$$\max\{\overline{F}_X(z), \overline{F}_Y(z)\} \leq \overline{F}_\Sigma(z) \leq 2\max\{\overline{F}_X(z/2), \overline{F}_Y(z/2)\}. \tag{4}$$

If those random variables are generalized Weibull-tail with tail parameters β_x and β_y, then there exist slowly-varying functions $l_1 = \max\{l_1^x, l_1^y\}$ and $l_2 = \min\{l_2^x, l_2^y\}$ with $l_1^x, l_2^x, l_1^y, l_2^y$ being slowly-varying function in the lower and upper bounds of generalized Weibull-tail X and Y respectively, such that

$$\mathrm{e}^{-z^\beta l_1(z)} \leq \max\{\overline{F}_X(z), \overline{F}_Y(z)\} \leq \mathrm{e}^{-z^\beta l_2(z)},$$

where $\beta = \min\{\beta_x, \beta_y\}$. Hence, Equation (4) is transformed into

$$\mathrm{e}^{-z^\beta l_1(z)} \leq \overline{F}_\Sigma(z) \leq \mathrm{e}^{-z^\beta \tilde{l}_2(z)},$$

where $\tilde{l}_2(z) = \frac{l_2(z/2)}{2^\beta}\left(1 - \frac{2^\beta \log 2}{z^\beta}\right)$ is slowly-varying. Thus, the survival function of the sum $X + Y$ is upper and lower-bounded by some Weibull-tail functions from family $\mathrm{WT}(\beta)$ where $\beta = \min\{\beta_x, \beta_y\}$. □

Theorem 2.2 (Product of generalized Weibull-tail RVs)**.** *The product of two independent non-negative generalized Weibull-tail random variables X and Y with tail parameters β_x and β_y is generalized Weibull-tail with tail parameter β such that $\frac{1}{\beta} = \frac{1}{\beta_x} + \frac{1}{\beta_y}$.*

Proof. Consider two independent non-negative generalized Weibull-tail random variables with tail parameters $\beta_x = \frac{1}{\theta_x}$ and $\beta_y = \frac{1}{\theta_y}$, $X \sim \mathrm{GWT}\left(\frac{1}{\theta_x}\right)$ and $Y \sim \mathrm{GWT}\left(\frac{1}{\theta_y}\right)$.

We want to prove that for some slowly-varying functions l_1 and l_2 the following holds for the product survival function :

$$\mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_1(z)} \leq \overline{F}_{XY}(z) = \mathbb{P}(XY \geq z) \leq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_2(z)}. \tag{5}$$

1. *Upper bound.* Firstly, notice that from the concavity of the logarithm for any $u, v > 0$ and $p \in (0,1)$ we have $\ln(pu + (1-p)v) \geq p\ln u + (1-p)\ln v$. Then $pu + (1-p)v \geq u^p v^{1-p}$. The change of variables $x = u^p$, $y = v^{1-p}$ implies

$$px^{1/p} + (1-p)y^{1/(1-p)} \geq xy. \tag{6}$$

From Equation (6), the upper bound of the product tail is

$$\mathbb{P}(XY \geq z) \leq \mathbb{P}\left(pX^{1/p} + (1-p)Y^{1/(1-p)} \geq z\right).$$

Lemma 2.1 implies that $pX^{1/p} \sim \mathrm{GWT}\left(\frac{p}{\theta_x}\right)$ and $(1-p)Y^{1/1-p} \sim \mathrm{GWT}\left(\frac{1-p}{\theta_y}\right)$. Taking $p = \frac{\theta_x}{\theta_x+\theta_y}$ and $1-p = \frac{\theta_y}{\theta_x+\theta_y}$, on the right-hand side we have a sum of two

independent generalized Weibull-tail random variables with tail parameter $\frac{1}{\theta_x+\theta_y}$. According to Theorem 2.1, this sum is generalized Weibull-tail with the same tail parameter $\frac{1}{\theta_x+\theta_y}$. It means that there exists slowly-varying function l_2 such that the tail of product XY is upper-bounded by

$$\mathbb{P}(XY \geq z) \leq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_2(z)}. \tag{7}$$

2. *Lower bound.* For independent non-negative X and Y we have

$$\mathbb{P}(XY \geq z) \geq \mathbb{P}\left(X \geq z^{\frac{\theta_x}{\theta_x+\theta_y}}\right) \mathbb{P}\left(Y \geq z^{\frac{\theta_y}{\theta_x+\theta_y}}\right).$$

Since X and Y are generalized Weibull-tail, it implies that

$$\mathbb{P}(XY \geq z) \geq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_1(z)}, \tag{8}$$

where $l_1(z) = l_1^x(z^{\theta_x/(\theta_x+\theta_y)}) + l_1^y(z^{\theta_y/(\theta_x+\theta_y)})$ with l_1^x and l_1^y being slowly-varying functions in the lower bounds of generalized Weibull-tail X and Y, is slowly-varying.

Combining together Equations (7) and (8) with Definition 2.1 implies the statement of the theorem. □

Since a Weibull-tail random variable is a particular case of a generalized Weibull-tail random variable, a sum or a product of Weibull-tail random variables will also give a generalized Weibull-tail random variable.

Example 2.1 (Sum of Gaussian and exponential RVs)**.** A convolution of Gaussian $\mathcal{N}(0, \sigma^2)$ and exponential $\mathrm{Exp}(\lambda)$ distributions can be written in the following form: $f_\Sigma(z) = \mathrm{e}^{-\lambda z + \frac{\lambda^2\sigma^2}{2} + \log\frac{\lambda}{2} + \log\left(1+\mathrm{erf}\left(\frac{z-\lambda\sigma^2}{\sqrt{2}\sigma}\right)\right)}$, where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-y^2} dy$ is the error function. Since $\mathrm{erf}\left(\frac{z-\lambda\sigma^2}{\sqrt{2}\sigma}\right) \to 1$ when $z \to \infty$, for big enough z the convolution $f_\Sigma(z) \sim \mathrm{e}^{-\lambda z + K}$ with $K > 0$. Then, there exists a slowly-varying function l such that $f_\Sigma(z) \sim \mathrm{e}^{-zl(z)}$. It means that the convolution of Gaussian and exponential distributions is WT(1), like the exponential one which is heavier among exponential and Gaussian.

3 Conclusion and discussion

We introduce a notion of a generalized Weibull-tail distribution, an extended version of Weibull-tail distribution with a property of stability. We showed that a sum of generalized Weibull-tail distributions (including dependent) is generalized Weibull-tail with tail parameter equal to a tail parameter of the heaviest distribution among them.

Here we considered only non-negative distributions, i.e. the *right* tail. The theory might be generalized to distributions on $\mathbb{R}$ by taking into account the *left* tails. In that case one can include asymmetric distributions where left and right tails have different tail parameters.

In application to Bayesian neural networks, Vladimirova et al. (2019) proved that hidden units, or neurons, follow a sub-Weibull distribution (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2018) where the tail parameter depends on depth. Future work will study the applicability of Weibull-tail properties to Bayesian neural networks.

References

Ahmad, Z., Mahmoudi, E., Hamedani, G., and Kharazmi, O. (2020). "New methods to define heavy-tailed distributions with applications to insurance data." *Journal of Taibah University for Science*, 14(1), 359–382.

Gardes, L. and Girard, S. (2016). "On the estimation of the functional Weibull tail-coefficient." *Journal of Multivariate Analysis*, 146, 29–45.

Gardes, L., Girard, S., and Guillou, A. (2011). "Weibull tail-distributions revisited: a new look at some tail estimators." *Journal of Statistical Planning and Inference*, 141(1), 429–444.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). "Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression." *arXiv preprint arXiv:1804.02605*.

Rachev, S. T. (2003). *Handbook of heavy-tailed distributions in finance: handbooks in finance, Book 1*. Elsevier.

Strupczewski, W. G., Kochanek, K., Markiewicz, I., Bogdanowicz, E., Weglarczyk, S., and Singh, V. P. (2011). "On the tails of distributions of annual peak flow." *Hydrology Research*, 42(2-3), 171–192.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. (2020). "Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions." *Stat*, 9(1), e318.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). "Understanding priors in Bayesian neural networks at the unit level." In *International Conference on Machine Learning*, 6458–6467. PMLR.

An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS

Thanh-Huan Vo [1], Guillaume Chauvet [2], André Happe [3], Emmanuel Oger [4], Stéphane Paquelet [5] & Valérie Garès [6]

[1] *INSA (IRMAR) and IRT b-com, Rennes, France, E-mail: than-huan.vo@insa-rennes.fr*
[2] *ENSAI (IRMAR), Campus de Ker Lann, Bruz, France, E-mail: guillaume.chauvet@ensai.fr*
[3] *EA 7449 REPERES, France, E-mail: andre.happe@chu-brest.fr*
[4] *EA 7449 REPERES, France, E-mail: emmanuel.oger@univ-rennes1.fr*
[5] *IRT b-com - Institut de Recherche Technologique b-com, France, E-mail: stephane.paquelet@b-com.com*
[6] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, E-mail: valerie.gares@insa-rennes.fr*

Résumé. Le couplage probabiliste d'enregistrements est un processus de combinaison de données provenant de différentes sources, lorsque ces données se réfèrent à des entités communes et que les informations d'identification ne sont pas disponibles. Fellegi et Sunter ont proposé un cadre de couplage probabiliste d'enregistrements quand les informations d'identification ne sont pas disponibles. Cependant, cette méthode n'utilise qu'une comparaison binaire entre les variables d'appariement. Dans nos travaux, nous proposons une extension de ce modèle notamment lorsque les données d'appariement contiennent différents types de variables (binaires, catégorielles et continues). Nous proposons un modèle de mélange de distributions discrètes pour gérer les variables d'appariement catégorielles avec une faible prévalence, et un modèle de mélange de distributions gamma gonflées en zéro pour gérer les variables d'appariement continues. Les estimations du maximum de vraisemblance pour les paramètres du modèle sont obtenues au moyen de l'algorithme "Expectation Conditional Maximization" (ECM). Grâce à une étude de simulation de Monte Carlo, nous évaluons à la fois l'estimation de la probabilité postérieure pour qu'une paire d'enregistrements soit une correspondance, et la qualité de prédiction des paires d'enregistrements appariés. Les premiers résultats de la simulation indiquent que les méthodes proposées donnent de bons résultats par rapport aux méthodes existantes. La prochaine étape consistera à appliquer la méthode proposée à un jeu de données réel, afin de trouver les patients correspondants dans les données des registres SNDS (Système National des Données de Santé) et GETBO (Groupe d'étude de la Thrombose de Bretagne Occidentale).

Mots-clés. Couplage d'enregistrements probabilistes, algorithme ECM, modèle de mélange, loi gamma gonflée en zéro

Abstract. Probabilistic record linkage is a process of combining data from different sources, when such data refer to common entities and that identifying information is not available. Fellegi and Sunter proposed a probabilistic record linkage framework that takes into account multiple non-identifying information but is limited to simple binary comparison between matching variables. In our work, we propose an extension of this model especially when matching data contains different types of variables (binary, categorical and continuous). We develop a model of mixture of discrete distribution for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables. The maximum likelihood estimates for model parameters are obtained by means of the Expectation Conditional Maximization (ECM) algorithm. Through a Monte Carlo simulation study, we evaluate both the posterior probability estimation for a record pair to be a match, and the prediction of matched record pairs. The first simulation results indicate that the proposed methods perform well as compared to existing methods. The next step will be to apply the proposed method to real datasets, which aim to find corresponding patients in SNDS (Système National des Données de Santé) and GETBO (Groupe d'étude de la Thrombose de Bretagne Occidentale) register data.

Keywords. Probabilistic record linkage, ECM algorithm, mixture model, hurdle gamma distribution

1 Introduction

Electronic health records have become more and more popular in medical fields, and the ability to exchange this information can help in providing better care for patients as well as richer sources for researchers. Record linkage is a process of combining data from different sources that refer to the same entity. The process is straightforward if each record contains a unique identifier such as Social Security Number. However, some large health databases may not contain such identifying information. Therefore, Fellegi and Sunter (1969) proposed a probabilistic record linkage framework that takes into account multiple non-identifying information such as names, and postal code.

Although this model is widely performed in many applications, when unique identifiers are unavailable or when data contain errors, its simple binary comparison has a limitation when some matching variables are binary and with a low prevalence (e.g. medical diagnoses, see Hejblum et al., 2019). Another limitation is that most probabilistic record linkage models only make use of simple binary or categorical comparison values even if the matching variables are continuous.

In this article, we propose a linkage model adapted from Fellegi and Sunter framework and which handle such cases. We aim at better taking into account the nature of matching

variables (e.g., low-prevalence binary, or continuous), so as to improve the performances of record linkage.

2 Probabilistic record linkage model

Consider two databases A and B containing n_A and n_B records respectively, and with elements in common. Following the terminology in Fellegi and Sunter (1969), each possible record pair $(X_{A,i}, X_{B,j})$ with

$$X_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}) \in A, i = 1, \ldots, n_A,$$
$$X_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}) \in B, j = 1, \ldots, n_B$$

either belongs to the set of true matched pairs noted by M, or to the set of true unmatched pairs noted by U.

The strategy begins by comparing K matching variables of all records $X_{A,i}$, with all records $X_{B,j}$ leading to $n_A \times n_B$ comparison vectors $\gamma_{ij} = \left\{\gamma^1_{ij}, \ldots, \gamma^k_{ij}, \ldots, \gamma^K_{ij}\right\}$, where $\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable which can be defined in different ways depending on the type of matching variables (see Christen, 2012). The most common way consists in a binary comparison, i.e.

$$\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j}) = \begin{cases} 1 & \text{if} \quad X^k_{A,i} = X^k_{B,j}, \\ 0 & \text{if} \quad X^k_{A,i} \neq X^k_{B,j}. \end{cases} \tag{1}$$

Because we assumed that each record pair belongs to one of two latent classes (the matched pairs M or the unmatched pairs U), the distribution of comparison vectors γ for each pair is assumed to follow a mixture model

$$\mathbb{P}(\gamma) = \mathbb{P}(\gamma|M)\mathbb{P}(\gamma \in M) + \mathbb{P}(\gamma|U)\left[1 - \mathbb{P}(\gamma \in M)\right]. \tag{2}$$

Once all the parameters of the model are estimated, the record pairs may be ordered and classified into matches, non-matches or possible matches based on either matching weights $\frac{\mathbb{P}(\gamma_{ij}|M)}{\mathbb{P}(\gamma_{ij}|U)}$ or posterior probabilities of matching $q_{ij} \equiv \mathbb{P}(M|\gamma_{ij}) = \frac{\mathbb{P}(\gamma_{ij}|M)p}{\mathbb{P}(\gamma_{ij}|M)p + \mathbb{P}(\gamma_{ij}|U)(1-p)}$. Although the matching scores and the posterior probabilities produce the same ordering for record pairs (Larsen and Rubin, 2001), the posterior probabilities are preferable in our application because they may be useful for further analyses (Lahiri and Larsen, 2005).

3 An extension of the Fellegi-Sunter model

In this article, we aim at developing the Fellegi-Sunter model by making better use of low prevalence categorical matching variables and of continuous variables.

3.1 Comparison approaches

Let X^k be a categorical matching variable taking L different values, which means that the comparison function for this variable may take L^2 values. For example, a comparison of a binary matching variable may lead to four possible realizations and a comparison function can be defined as follows

$$h^k(0,0) = 0,\ h^k(0,1) = 1,\ h^k(1,0) = 2, \quad \text{and} \quad h^k(1,1) = 3. \tag{3}$$

Because the agreement on the low prevalence value is much more informative than the agreement on the others, our comparison approach aims at using this information while the simple binary comparison (1) method does not distinguish them. Hejblum et al. (2018) propose a Bayesian record linkage framework making use of a similar idea, and which is efficient in case of a large number of low-prevalence binary matching variables. However, their model is designed for binary variables only, while our comparison approach can be combined with other types of matching variables (e.g., continuous).

If the number of matching variables and/or the number of categories is large, the number of parameters to be estimated is L^2-1, which may be too large in practice. This number may be reduced by assigning a same comparison value for the agreement/disagreement of categories which have roughly a same proportion. For instance, we may reduce the comparison values given in (3) as

$$h^k(0,0) = 0,\ h^k(0,1) = h^k(1,0) = 1, \quad \text{and} \quad h^k(1,1) = 2. \tag{4}$$

Now, let us consider the case of a continuous variable X^k. For example date variables (e.g., admission to the hospital, or medical act) are common in medical datasets. They may be seen as continuous counting variables, by converting each date into a duration from a specified origin. Even if an individual is present in both datasets, a lag between dates is likely to appear. The simple binary comparison is therefore not appropriate. In this article, we propose to define $\gamma_{ij}^k = d(X_{A,i}^k, X_{B,j}^k)$, where d is a predefined distance, which can naturally take into account the time lags.

In summary, the comparison vectors in our model can include both categorical and continuous comparison values.

3.2 Estimation of parameters

Let

$$\gamma_{ij} \;=\; \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}, \gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}\right) \tag{5}$$

be a mixed type comparison vector which includes K_1 categorical comparison values $\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}$ and K_2 continuous distances $\gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}$. Following the Fellegi-Sunter framework, these comparison vectors are assumed to follow the mixture model (2).

Under the conditional independence assumption between fields of the comparison vector (Winkler, 2000), we have

$$\mathbb{P}(\gamma_{ij}|M) = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M) \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M) \tag{6}$$

$$\mathbb{P}(\gamma_{ij}|U) = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U) \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U) \tag{7}$$

For the first term in the right-hand side involving K_1 categorical comparison values of the comparison vector γ_{ij}, we define

$$m_s^k = \mathbb{P}(\gamma_{ij}^k = s|M), \sum_{s\in S^k} m_s^k = 1, \text{ and } u_s^k = \mathbb{P}(\gamma_{ij}^k = s|U), \sum_{s\in S^k} u_s^k = 1$$

for S^k is the set of all possible categorical comparison values for the k^{th} variable. Then we have

$$\mathbb{P}(\gamma_{ij}^k|M) = \prod_{s\in S^k} (m_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}}, \quad \text{and} \quad \mathbb{P}(\gamma_{ij}^k|U) = \prod_{s\in S^k} (U_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}} \quad \text{for} \quad k = 1, \ldots, K_1.$$

For the second part in the right-hand side of equations (6) and (7) which involves K_2 continuous values of the comparison vector γ, we define

$$\begin{aligned} \mathbb{P}(\gamma_{ij}^k|M) &\sim f^k(\phi_M^k), \\ \text{and} \quad \mathbb{P}(\gamma_{ij}^k|U) &\sim f^k(\phi_U^k) \end{aligned}$$

for $k = K_1 + 1, \ldots, K_2$. The distribution f^k needs to be postulated, depending on the characteristics of the continuous matching variables and the chosen distance. In our simulation studies, we model f^k by means of a hurdle Gamma distribution. To find the maximum likelihood estimates for parameters, we apply the the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993).

4 Simulation studies

For ease of interpretation, our proposed approaches are evaluated and compared to existing approaches over scenarios for binary and continuous variables separately. We will consider two databases A and B containing n_A and n_B individuals and K matching variables. We assume that there is no duplicate in both databases and that all individuals in B have corresponding individuals in A. To be realistic, we also introduced errors for data in B and the distribution of error depends on type of each matching variable.

When there are only binary matching variables, we compare the method proposed by Hejblum et al. (2019) to the Fellegi-Sunter model with different comparison methods (1), (4) and (3). When there are only continuous matching variables, we compare the Fellegi-Sunter model using hurdle gamma distribution for continuous comparison values to this model using discrete distribution for categorical comparison values.

The record linkage procedures are evaluated by means of two common criteria for an imbalance classification problem which are True positive rate (Sensitivity or Recall) and Positive predictive value (Precision). From the record linkage results of different considered cases, there is a significant improvement of the Fellegi-Sunter model with our proposed comparison approaches compared to the model with simple binary comparison.

5 Application

The SNDS database is the French national health database that includes all health insurance and hospital data. It is therefore of major interest for research. The GETBO is a registry database that collects information of venous thromboembolism cases in Brest, France. These databases have common information on demographic data such as month and year of birth, gender and some medical acts. The objective is to link the registry data to SNDS at the patient level, when no common individual identifier is available. Our proposed approach will be performed on these databases.

Bibliography

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection, *Springer Publishing Company, Incorporated*.
Fellegi, I. and Sunter, A. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64, pp. 1183-1210.
Hejblum, B., Weber, G., Liao, K., Palmer, N., Churchill, S., Shadick, N., Szolovits, P., Murphy, S., Kohane, I., and Cai,T. (2019). Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes, *Scientific Data*.
Lahiri, P. and Larsen, M.B. (2005). Regression analysis with linked data, *Journal of the American Statistical Association*, 100 (469), pp. 222-230.
Larsen, M.B. and Rubin, D.B. (2001). Iterative automated record linkage using mixture models, *Journal of the American Statistical Association*, 96 (453), pp. 32-41.
Meng, X. and Rubin, D. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80(2), pp. 267-278.
Winkler, W.E. (2000). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *U.S. Bureau of the Census*.

Consistency of the k-nearest neighbor rule of classification for spatial training data

Ahmad YOUNSO [1] & Nour AZHARI [2]

[1] *Damascus Uniersity, Department of mathematical statistics, Damascus, Syria.*
E-mail: ahyounso@yahoo.fr
[2] *Tishreen University, Department of Mathematics, Lattakia, Syria.*
E-mail: n.azhari@tishreen.edu.sy

Résumé. Le but de ce travail est d'étudier la règle de classification du plus proche voisin pour des données spatialement dépendantes. Certaines conditions de mélange spatial sont considérées, et dans de telles structures spatiales, la règle du plus proche voisin est suggérée pour classer des données spatiales sous des hypothèses modérées. La consistance et la consistance forte de la règle sont établies. Les résultats de ce travail étendent des précédents résultats obtenus dans le i.i.d. cas au cas spatial.

Mots-clés. Règle de Bayes, données d'apprentissage, règle du plus proche voisin, condition de mélange, consistance.

Abstract. The purpose of this work is to investigate the k-nearest neighbor classification rule for spatially dependent data. Some spatial mixing conditions are considered, and under such spatial structures, the well known k-nearest neighbor rule is suggested to classify spatial data. We established consistency and strong consistency of the classifier under mild assumptions. The main results of this work extend the consistency result in the i.i.d. case to the spatial case.

Keywords. Bayes rule, training data, k-nearest neighbor rule, mixing condition, consistency.

1 k-nearest neighbor rule for spatial training data

Analysis of spatial data arises in various areas of research including agricultural field trials, astronomy, econometrics, epidemiology, environmental science, geology, hydrology, image analysis, meteorology, ecology, oceanography and many others in which the data of interest are collected across space. One of the most fundamental issues in spatial analysis is classification and pattern recognition. For example, in remote sensing technology or digital geography information, we need somehow to classify spatial data into patterns or images into types. The aim of the present work is to investigate whether the classical k-nearest neighbor classifier can be extended to classify spatial data. The use of the k-nearest neighbor (k-NN) method in the spatial case is due to Tran (1993) for density estimation.

The real interest in the k-NN method comes from the nature of the smoothing parameter. Indeed, in the traditional kernel method, the smoothing parameter is the bandwidth, which is a real positive number. Here, the number of neighbors k is the smoothing parameter and it takes its values in a discrete set. The main difficulties with the kernel method appear when data are sparse; choosing the number of neighbors allows to avoid this problem and is adapted to the concentration of the data. Consistency of kernel-based rules on temporally or spatially dependent data has recently been investigated by Younso (2017, 2018, 2019, 2020) in finite and infinite-dimensional space. In the present work, we will establish the (strong) consistency of the k-nearest neighbor classifier based on spatially dependent training data. Let $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ be a strictly stationary random field defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^d \times \{0,1\}$. In the problem of classification, for each $\mathbf{i} \in \mathbb{Z}^N$, $X_{\mathbf{i}}$ is a vector of features and $Y_{\mathbf{i}}$ is the label (class) of $X_{\mathbf{i}}$. A point $\mathbf{i} = (i_1, ..., i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, ..., n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_l \leq n_l, \ \forall l = 1, ..., N\}$. We will write $\mathbf{n} \to \infty$ if $\min_{1\leq l\leq N} n_l \to \infty$. Define $\hat{\mathbf{n}} = n_1 \times ... \times n_N = \text{card}\,(\mathcal{I}_{\mathbf{n}})$. In a new site $\mathbf{j}$, one wishes to predict the label $Y_{\mathbf{j}}$ of an observation $X_{\mathbf{j}}$. The pair $(X_{\mathbf{j}}, Y_{\mathbf{j}})$ may be described by μ, the probability measure for $X_{\mathbf{j}}$, and $\eta(x) = \mathbb{E}\,(Y_{\mathbf{j}}/X_{\mathbf{j}} = x)$, the regression of $Y_{\mathbf{j}}$ on $X_{\mathbf{j}} = x$. Assume that for each $\mathbf{i} \in \mathbb{Z}^N$, $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has the same distribution as the pair (X, Y). We create a classifier $g : \mathbb{R}^d \longrightarrow \{0,1\}$ mapping $X_{\mathbf{j}}$ into the predicted label of $X_{\mathbf{j}}$. The error rate, or risk, of a rule g is $L(g) = \mathbb{P}\{g(X_{\mathbf{j}}) \neq Y_{\mathbf{j}}\}$. This is minimized by the rule

$$g^*(x) = \begin{cases} 0 \text{ if } \mathbb{P}\{Y_{\mathbf{j}} = 0 | X_{\mathbf{j}} = x\} \geq \mathbb{P}\{Y_{\mathbf{j}} = 1 | X_{\mathbf{j}} = x\} \\ 1 \text{ otherwise,} \end{cases}$$

whose error rate $L^* = L(g^*)$ is called the Bayes risk and g^* is called the Bayes rule (see Devroye, Györfi and Lugosi (1996)). This optimal rule depends on the distribution of $(X_{\mathbf{j}}, Y_{\mathbf{j}})$ which is generally unknown. We use the data $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$ to construct a classifier $g_{\mathbf{n}}(x)$. The set $D_{\mathbf{n}}$ is called training sample. The spatial version of the classical k-NN rule is given by

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \displaystyle\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} w_{\mathbf{ni}} Y_{\mathbf{i}} \leq 1/2 \\ 1 \text{ otherwise,} \end{cases} \tag{1}$$

where $w_{\mathbf{ni}} = w_{\mathbf{ni}}(x; D_{\mathbf{n}})$ is $1/k$ if $X_{\mathbf{i}}$ is one of the k-nearest neighbor of x in $D_{\mathbf{n}}$ and $w_{\mathbf{ni}}$ is zero otherwise with $k = k(\mathbf{n})$ is a sequence of positive integers satisfying

$$k \longrightarrow \infty \quad \text{and} \quad k/\hat{\mathbf{n}} \longrightarrow 0 \ as \ \mathbf{n} \to \infty. \tag{2}$$

According to (1), to classify a new obsevation $X_{\mathbf{j}} = x$, the k-NN rule adopts the majority case of its k numbers of neighbors as the class suggested. To break tie, if $X_{\mathbf{i}}$ and $X_{\mathbf{i}'}$ are equidistant from x, one chooses $X_{\mathbf{i}}$ if $\|\mathbf{i} - \mathbf{j}\| < \|\mathbf{i}' - \mathbf{j}\|$ with $\|.\|$ denotes the Euclidean norm. We assume that X has a density f, so that we can avoid messy technicalities

necessary to handle distance ties. Denote $L_{\mathbf{n}} = \mathbb{P}\left(Y_{\mathbf{j}} \neq g_{\mathbf{n}}(X_{\mathbf{j}})\right)$ the probability of error by $g_{\mathbf{n}}(x)$. The best we can expect from $g_{\mathbf{n}}(x)$ is to achieve the Bayes risk. The classifier $g_{\mathbf{n}}(x)$ is called consistent if $\mathbb{E}L_{\mathbf{n}} \longrightarrow L^*$ as $\mathbf{n} \to \infty$ and it is called strongly consistent if $L_{\mathbf{n}} \longrightarrow L^*$ as $\mathbf{n} \to \infty$ with probability one. Under l'assupmtion (2), the consistency of the k-NN rule when the training dataset is i.i.d. was proved by Stone (1977). In this work, we investigate both the consistency and strong consistency of $g_{\mathbf{n}}$ when the training dataset is spatially dependent.

1.1 Mixing conditions

Let $\mathcal{A}$ and $\mathcal{C}$ be two sub σ-algebras of $\mathcal{F}$. The α-mixing coefficient between $\mathcal{A}$ and $\mathcal{C}$ is defined by

$$\alpha = \alpha(\mathcal{A}, \mathcal{C}) = \sup_{A \in \mathcal{A}, C \in \mathcal{C}} |\mathbb{P}(A \cap C) - \mathbb{P}(A)\mathbb{P}(C)|$$

and the β-mixing coefficient is defined by

$$\beta = \beta(\mathcal{A}, \mathcal{C}) = \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{P}(A|\mathcal{C}) - \mathbb{P}(A)]|.$$

Let $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^N}$ be a random field on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in some space $(\Omega', \mathcal{F}')$. For any $E, E' \subset \mathbb{Z}^N$ with finite cardinals, we denote by $\mathcal{B}(E)$ and $\mathcal{B}(E')$ the Borel σ-algebras generated by $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in E}$ and $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in E'}$ respectively. The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^N}$ is said to be α-mixing (strongly mixing) if

$$\alpha(t) = \sup_{\mathrm{dist}(E,E') \geq t} \alpha\big(\mathcal{B}(E), \mathcal{B}(E')\big) \downarrow 0 \text{ as } t \to \infty,$$

where

$$\mathrm{dist}(E, E') = \inf_{\mathbf{i} \in E, \mathbf{j} \in E'} \|\mathbf{i} - \mathbf{j}\|$$

and $\|.\|$ denotes the Euclidean norm. The above α-mixing condition may be satisfied by many spatial models and examples can be found in Neaderhouser (1980) and Rosenblatt (1985). The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^N}$ is said to be β-mixing (absolutely regular) if

$$\beta(t) = \sup_{\mathrm{dist}(E,E') \geq t} \beta\big(\mathcal{B}(E), \mathcal{B}(E')\big) \downarrow 0 \text{ as } t \to \infty.$$

The two mixing coefficients α and β are related by the inequality $2\alpha \leq \beta$ (see Rio (2000)). Consequently, any β-mixing random field is α-mixing one.

1.2 Main results

In the following theorem, we investigate consistency of the k-nearest neighbor rule.

Theorem 1. *Suppose that D_n are observations of α-mixing random field such that $\alpha(t) = O(t^{-\theta})$ with $\theta > N$. Suppose in addition that (2) is satisfied and that as $\mathbf{n} \to \infty$,*

$$k/\sqrt{\hat{\mathbf{n}}} \longrightarrow \infty. \tag{3}$$

Then, as $\mathbf{n} \to \infty$,

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^*.$$

Theorem 1 extends the consistency of the k-nearest neighbor established by Devroye, Györfi and Lugosi (1996) to the spatial case. Observe that the additional condition (3) is weaker than that used by (Bosq and Lecoutre (1987), Theorem II.3, p. 234) in the i.i.d. regression estimate case. In the following theorem, we investigate strong consistency of the k-nearest neighbor rule.

Theorem 2. *Suppose that D_n are observations of strictly stationary β-mixing random field such that $\beta(t) = O(t^{-\theta})$ with $\theta > 2N$ and that (2) and (3) are satisfied. Suppose in addition that there is an integer $p = p(\mathbf{n}) \in [1, \min_{1\leq l\leq N} n_l/2]$ such that as $\mathbf{n} \to \infty$, $p \to \infty$,*

$$\hat{\mathbf{n}}/(p \log \hat{\mathbf{n}}) \longrightarrow \infty \tag{4}$$

and

$$\sum_{\hat{\mathbf{n}}\in(\mathbb{N}^*)^N} k^{-1}\hat{\mathbf{n}}\beta(p) < \infty. \tag{5}$$

Then, as $\mathbf{n} \to \infty$,

$$L_{\mathbf{n}} \longrightarrow L^* \quad \text{with probability one.}$$

Observe that if we take for example $p = \hat{\mathbf{n}}^{\gamma}$, the additional (4) and (5) are satisfied for some $2/\theta < \gamma < 1$.

1.3 Selection of the best nearest neighbor

The choice of the number of neighbors k is an essential point of the k-NN method. It is desirable for k to be odd to make ties less likely. We use the cross-validation criteria to approximate the best k. With this criteria, the dataset will be splitted into three parts: training data, cross-validation data, and test data. We use training data for finding nearest neighbors (integer numbers being greater than $\sqrt{\hat{\mathbf{n}}}$ according to (3)), we use cross-validation data to find the best value of k and finally we test our model on totally unseen test data. For sparse data sets, leave-one-out (LOO or LOOCV) may need to be used.

Bibliographie

Bosq, D. and Lecoutre,J.P. *Théorie de l'estimation fonctionnelle.* Eco- nomica, Paris, 1987. Devroye, L., Györfi, L. and Lugosi, G. *A probabilitic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

Neaderhouser, C. C. Convergence of block spins defined on random fields. *J. Statist. Phys.*, 22:673-684, 1980.

Rio, E. *Théorie asymptotique des processus aléatoires faiblement dépendents.* Mathématiques et Applications . Spriner, Berlin, 2000.

Rosenblatt, M. *Stationary sequences and random fields.* Birkhauser, Boston, 1985.

Stone, C. J. Consistent Nonparametric Regression. *Annals of Statistics.* 4:595-620, 1977.

Tran, L. T. and Yakowitz, S. Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, 44:23-46, 1993.

Younso, A. On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters*, 131:64-71, 2017.

Younso, A. On the consistency of kernel classification rule for functional random field. *Journal de la Société Française de Statistique*, 159:68-87, 2018.

Younso, A., Kanaya, Z. and Azhari, N. Strong consistency of a kernel- based rule for spatially dependent data. *Arab Journal of Mathematical Sciences*, 26:211-225, 2019.

Younso, A. Nonparametric discrimination of areal functional data. *Brazilian Journal of Probability and Statistics*, 34:12-126, 2020.

Estimation de la fonction de variance par agrégation de type sélection modèle

Ahmed ZAOUI

Laboratoire LAMA, Université Gustave Eiffel,
Ahmed.Zaoui@univ-eiffel.fr

Résumé. Dans ce travail, nous nous intéressons à l'estimation de la fonction de variance en régression par agrégation de type sélection modèle (`MS`). Le but de la procédure `MS` est de sélectionner le meilleur estimateur parmi un ensemble de prédicteurs. Le prédicteur sélectionné est alors appelé `MS`-estimateur. La construction de `MS`-estimateur repose sur une procédure en deux étapes. Dans une première étape, à partir d'un premier échantillon, nous construisons des estimateurs de la fonction de variance par la méthode basée sur les erreurs résiduelles. Dans une deuxième étape, nous les agrégeons à l'aide d'un deuxième échantillon. Nous établissons la consistance de `MS`-estimateur vis-à-vis du risque L_2 et illustrons ses performances numériques sur simulations.

Mots-clés. Agrégation, Régression, Méthode basée sur les erreurs résiduelles.

Abstract. In this work, we focus on the variance function estimation in regression by model selection aggregation `MS`. The aim of the `MS` procedure is to select the best estimator from a set of predictors. The selected predictor is then called `MS`-estimator. The construction of `MS`-estimator relies on a two-step procedure. In the first step, from a first sample, we construct estimators of the variance function by the residual-based method. In the second step, we aggregate them using a second sample. We establish the consistency of `MS`-estimator with respect to the L_2 risk and illustrate its numerical performances on simulations.

Keywords. Aggregation, Regression, Residual-based method.

1 Introduction

Nous introduisons tout d'abord le modèle de régression. Dans ce cadre, une donnée observée est de la forme (X, Y) où $X \in \mathbb{R}^d$ est la variables explicative et $Y \in \mathbb{R}$ est la variable à prédire associée à l'entrée X telle que

$$Y = f^*(X) + \zeta,$$

où ζ est la variable de bruit satisfaisant $\mathbb{E}[\zeta|X] = 0$ et $\mathbb{E}[\zeta^2] < \infty$. Dans la suite, nous notons $f^*(x) = \mathbb{E}[Y|X = x]$ la fonction de régression et $\sigma^2(x) = \mathbb{E}[(Y - f^*(X))^2|X = x]$ la fonction de variance conditionnelle pour tout $x \in \mathbb{R}^d$.

L'estimation de la fonction de variance conditionnelle joue un rôle important en régression, notamment pour mesurer la volatilité ou le risque en finance (Anderson et al (1997)), ou encore pour la construction d'un intervalle de confiance pour la fonction de régression (Hart (1997)). Plus récemment, dans le cadre de la régression avec option rejet, (Denis et al (2020)) ont montré que le prédicteur optimal repose sur une seuillage de la fonction de variance. Dans ce travail, nous proposons une méthode d'agrégation pour estimer la fonction de variance.

Dans la littérature, de nombreuses méthodes sont proposées pour estimer la fonction de variance conditionnelle. Les deux méthodes les plus populaires sont la méthode directe et la méthode basée sur les erreurs résiduelles. La méthode directe (Härdle et al(1997)) repose sur une décomposition de la fonction de variance conditionnelle σ^2 qui est réécrite comme la différence des deux premiers moments conditionnels, $\sigma^2(X) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2$. Elle consiste à estimer séparément les deux termes du côté droit. **La méthode basée sur les erreurs résiduelles** consiste en deux étapes. Dans une première étape, nous construisons un estimateur $\hat{f}$ de la fonction de régression f^*. Dans une deuxième étape, un estimateur de σ^2 est obtenu en résolvant le problème de régression où la variable d'entrée est X et la variable à prédire est $(Y - \hat{f}(X))^2$. Pour plus de détails, nous renvoyons à Ruppert et al (1997), Fan et al (1998), Kulik et al (2011) et Denis et al (2020). Dans ce travail, nous nous concentrons sur la méthode basée sur les erreus résiduelles pour estimer la fonction de variance, car cette procédure fournit de bonnes garanties théoriques et numériques.

L'agrégation est une approche populaire en apprentissage statistique pour estimer f^* dans le modèle de régression. Pour plus de détails, nous renvoyons à Nemirovski (2000), Tsybakov (2003), Yang (2004) et Tsybakov (2014). Une méthode très utilisée en pratique est l'agrégation par sélection de modèle (`MS`). Étant donné un dictionnaire d'estimateurs de la fonction de régression, `MS` consiste, sur la base d'un échantillon d'apprentissage, à sélectionner au sein du dictionnaire le meilleur prédicteur. Dans ce travail, nous appliquons le principe de la méthode `MS` pour estimer la fonction de variance. À notre connaissance, ce travail est le premier à étendre la notion d'agrégation à l'estimation de σ^2.

Notations. Soit $p \geq 2$ un entier, $[p] := \{1, \cdots, p\}$. Soit N un entier, pour toute fonction $f : \mathbb{R}^d \to \mathbb{R}$, nous définissons la norme empirique $\|f\|_N^2 = \frac{1}{N} \sum_{i=1}^N |f(X_i)|^2$.

2 Agrégation de type sélection modèle

Cette section est dédiée à l'estimation de la fonction de variance par `MS` et à l'étude de la consistance de la procédure proposée. Nous rappelons que nous nous concentrons sur **la méthode basée sur les erreus résiduelles** pour estimer la fonction de variance.

2.1 Méthode

Dans cette section, nous décrivons l'algorithme d'estimation de la fonction de variance σ^2 en utilisant une méthode agrégation de type sélection modèle. L'estimateur résultant est appelé MS-estimateur. Nous introduisons d'abord deux échantillons d'apprentissage indépendants : $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \cdots, n\}$ et $\mathcal{D}_N = \{(X_i, Y_i), i = n+1, \cdots, n+N\}$ qui consistent en n et N copies indépendantes et identiquement distribuées de (X, Y). La méthode que nous proposons est en deux étapes. Dans la première étape, nous considérons M_1 estimateurs de la fonction de régression $\hat{f}_1, \cdots, \hat{f}_{M_1}$ basée sur $\mathcal{D}_n$. Ensuite, nous utilisons le deuxième échantillon $\mathcal{D}_N$ pour estimer f^* par MS : nous sélectionnons l'indice optimal $\hat{s}$

$$\hat{s} \in \underset{s\in[M_1]}{\operatorname{argmin}} \mathcal{R}_N(\hat{f}_s), \text{ avec } \mathcal{R}_N(\hat{f}_s) = \frac{1}{N}\sum_{i=1}^{N} |Y_i - \hat{f}_s(X_i)|^2 \ ,$$

et le MS-estimateur de la fonction de régression, noté $\hat{f}_{\texttt{MS}}$, est donné comme suit

$$\hat{f}_{\texttt{MS}} := \hat{f}_{\hat{s}}.$$

Dans une deuxième étape, étant donné l'estimateur $\hat{f}_{\texttt{MS}}$ ($\hat{f}_{\hat{s}}$) construit sur $\mathcal{D}_N$, nous construisons M_2 estimateurs de la fonction de variance σ^2, construits à partir de $\mathcal{D}_n$, par **la méthode basée sur les erreurs résiduelles**. Ces estimateurs sont notés $\hat{\sigma}^2_{\hat{s},1}, \cdots, \hat{\sigma}^2_{\hat{s},M_2}$. Enfin, sur la base de $\mathcal{D}_N$, nous sélectionnons l'indice optimal, noté $\hat{m}$, comme suit

$$\hat{m} \in \underset{m\in[M_2]}{\operatorname{argmin}} \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) \text{ où } \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{Z}_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$$

avec $\hat{Z}_i = \left(Y_i - \hat{f}_{\texttt{MS}}(X_i)\right)^2$. Par conséquent, le MS-estimateur de la fonction de variance, noté $\hat{\sigma}^2_{\texttt{MS}}$, est défini comme suit

$$\hat{\sigma}^2_{\texttt{MS}} := \hat{\sigma}^2_{\hat{s},\hat{m}}.$$

2.2 Résultat principal

Cette section est consacrée à l'étude du risque L_2 de $\hat{\sigma}^2_{\texttt{MS}}$. Soit $\mathcal{R}(\hat{f}_s) = \mathbb{E}\left[|Y - \hat{f}_s(X)|^2\right]$ le risque quadratique pour $\hat{f}_s$ pour tout $s \in [M_1]$. Nous définissons s^* comme suit

$$s^* \in \underset{s\in[M_1]}{\operatorname{argmin}} \mathcal{R}(\hat{f}_s)$$

Nous introduisons également les hypothèses suivantes :

Hypothèse 1. *Les fonctions f^* et σ^2 sont bornées.*

Hypothèse 2. *Pour tout $s \in [M_1]$ et tout $m \in [M_2]$, $\hat{f}_s$ et $\hat{\sigma}^2_{s,m}$ sont bornés.*

Hypothèse 3 (Hypothèse de séparabilité)**.** *Il existe $\delta_0 > 0$ telle que*

$$\delta^*(\mathcal{D}_n) = \min_{s \neq s^*} \{|\mathcal{R}(\hat{f}_s) - \mathcal{R}(\hat{f}_{s^*})|\} > \delta_0 \ .$$

Hypothèse 4. *Y est borné ou Y satisfait le modèle gaussien*

$$Y = f^*(X) + \sigma(X)\xi,$$

où $\xi \sim \mathcal{N}(0,1)$ est indépendante de X.

Ces hypothèses jouent un rôle crucial sur l'étude de la consistance de $\hat{\sigma}^2_{\texttt{MS}}$. Nous pouvons à présent établir notre résultat principal :

Théorème 1. *Soit $\hat{f}_{MS}$ et $\hat{\sigma}^2_{MS}$ les MS-prédicteurs de f^* et σ^2 respectivement. Sous les Hypothèses 1, 2, 3 et 4, il existe deux constantes absolues $C_1 > 0$ et $C_2 > 0$ telle que*

$$\mathbb{E}\left[|\hat{\sigma}^2_{MS}(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\min_{m \in [M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C_1\sqrt{\min_{s \in [M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|^2_N\right]} + C_2\phi^{MS}_N(M_1) \ , \quad (1)$$

où

$$\phi^{MS}_N(M_1) = \begin{cases} \left(\frac{\log(M_1)}{N}\right)^{1/4} & \text{si } Y \text{ est borné;} \\ \left(\frac{\log(M_1)}{N}\right)^{1/8} & \text{sinon.} \end{cases}$$

Théorème 1 donne une borne supérieure pour le risque L_2 de $\hat{\sigma}^2_{\texttt{MS}}$. Le premier terme dans le côté droit de l'équation (1) représente le biais de MS-estimateur $\hat{\sigma}^2_{\texttt{MS}}$ qui dépend de s^*, tandis que le deuxième est du à l'erreur d'estimation de la fonction de régression f^*. Le troisième terme est un terme de variance qui est d'ordre ; $(\log(M_1)/N)^{1/4}$ dans le cas où Y est borné et $(\log(M_1)/N)^{1/8}$ dans le cas où Y n'est pas borné. Cette vitesse lente est du au fait que l'estimation de la fonction de variance repose sur f^* que l'on doit également estimer.

3 Simulations

Dans cette section, nous étudions les performances numériques de MS-estimateur $\hat{\sigma}^2_{\texttt{MS}}$. La construction de $\hat{\sigma}^2_{\texttt{MS}}$ est détaillée en section 2.1. Nous introduisons deux ensembles $\mathcal{F} = \{\hat{f}_s\}_{s=1}^6$ et $\Sigma = \left\{\hat{\sigma}^2_{\hat{s},m}\right\}_{m=1}^6$ (dont la construction repose sur $\hat{f}_{\texttt{MS}}$) qui contiennent six estimateurs construits à partir des algorithmes des forêt aléatoire (rf), des k-plus proches voisins (K-PPV), du Lasso, et des machines à vecteurs de support (svm) basés

sur les noyaux de type base radiale, polynomiale et sigmoïde. Pour les algorithmes svm et rf, nous utilisons respectivement les packages `R`, `e1071` et `randomForest` avec **des paramètres par défaut**. Pour K-PPV et Lasso, nous utilisons le package `FNN` et `glmnet` respectivement. La sélection de l'entier k et du coefficient de pénalité λ est effectuée par validation croisée. Enfin, les performances de l'estimateur $\hat{\sigma}^2_{\texttt{MS}}$ sont évaluées comme suit. On répète indépendamment 100 fois les étapes suivantes :

(i) On simule trois ensembles de données $\mathcal{D}_n$, $\mathcal{D}_N$ et $\mathcal{D}_T$ avec $n, N \in \{100, 1000\}$, et $T = 1000$.

(ii) À partir de $\mathcal{D}_n$, nous construisons les estimateurs constituant $\mathcal{F}$, puis à partir de $\mathcal{D}_N$, nous calculons $\hat{f}_{\texttt{MS}}$. Ensuite à partir de $\mathcal{D}_n$ et $\hat{f}_{\texttt{MS}}$, nous calculons les estimateurs constituant Σ et puis nous calculons $\hat{\sigma}^2_{\texttt{MS}}$ sur $\mathcal{D}_N$.

(iii) À partir de $\mathcal{D}_n \cup \mathcal{D}_N$: dans un premier temps, nous calculons les estimateurs constituant $\mathcal{F}$; dans un deuxième temps, pour chaque estimateur $\hat{f}_s$ de $\mathcal{F}$ nous calculons les estimateurs $\{\hat{\sigma}^2_{s,m}\}_{1\leq m\leq 6}$ pour les six procédures.

(iv) Enfin, sur $\mathcal{D}_T$, nous calculons l'erreur empirique L^2 ($\widehat{\text{Err}}$) de l'agrégat $\hat{\sigma}^2_{\texttt{MS}}$ et de tous les estimateurs de la fonction de variance σ^2 obtenus à l'étape (iii). Nous choisissons le meilleur estimateur parmi eux que l'on l'appelle $\hat{\sigma}^2_{\texttt{Best}}$.

À partir de ces estimations, nous calculons la moyenne et l'écart-type de $\widehat{\text{Err}}$. Pour notre étude numérique, nous considérons le modèle suivant

$$Y = f^*(X) + \sigma(X)\xi,$$

où $\xi \sim \mathcal{N}(0,1)$ indépendant à X. Nous considérons deux modèles

- Modèle 1 : soit X une distribution uniforme sur $[0,1]^3$ telle que
  1. $f^*(X) = \cos(2X_1) + X_2$
  2. $\sigma^2(X) = \frac{1}{4}\left(0.1 + \exp(-7(X_1 - 0.2)^2) + \exp(-10(X_2 - 0.8)^2 + \exp(-20(X_3 - 0.9)^2\right)$
- Modèle 2 : soit $X = (X_1, \cdots, X_{10})$ une distribution uniforme sur $[0,1]^{10}$ telle que
  1. $f^*(X) = 0.01 + X_1 + X_2 + X3 + X_{10}$
  2. $\sigma^2(X) = \left(0.9 + (X_1(1 - X_2))^{\frac{1}{2}} \sin\left(\frac{2.1\pi}{X_3 + 0.05}\right) + 0.1\exp\left(-550(X_7 - 0.8)^2\right)\right)^2$.

Le modèle 1 est un modèle multivarié dans lequel la fonction de variance prend des valeurs relativement modérées ($0.030 \leq \sigma^2(X) < 0.765$). Par contre pour le modèle 2 qui est aussi un modèle multivarié, la fonction de variance peut prendre de grandes valeurs ($\sigma^2(X) \in]0, 4.432]$, avec 41.5% des valeurs sont supérieures à 1). Le modèle 2 est donc un modèle où l'estimation de la fonction de variance est difficile.

Le résultats sont donnés dans le tableau 1. Nous faisons deux observations. Premièrement, lorsque n et N sont assez grands, le `MS`-estimateur $\hat{\sigma}^2_{\texttt{MS}}$ a des performances similaires à celles du meilleur estimateur $\hat{\sigma}^2_{\texttt{Best}}$ qui est construit à partir $\mathcal{D}_n \cup \mathcal{D}_N$. Deuxièmement, nous remarquons dans le tableau 1, que pour le Modèle 1, l'erreur empirique $\widehat{\text{Err}}$ de $\hat{\sigma}^2_{\texttt{MS}}$ converge

TABLE 1 – Moyenne et écart-type de l'erreur empirique L^2 des deux estimateurs.

	$n = N = 100$		$n = 100,\ N = 1000$		$n = 1000,\ N = 100$		$n = N = 1000$	
	MS	Best	MS	Best	MS	Best	MS	Best
Model	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
Model 1	0.03 (0.023)	0.01 (0.004)	0.02 (0.013)	0.01 (0.001)	0.02 (0.022)	0.01 (0.001)	0.006 (0.002)	0.004 (0.000)
Model 2	0.61 (0.167)	0.45 (0.031)	0.53 (0.112)	0.41 (0.029)	0.46 (0.120)	0.40 (0.028)	0.43 (0.034)	0.37 (0.033)

plus rapidement vers $\hat{\sigma}^2_{\texttt{Best}}$ que pour le Modèle 2. En effet, nous avons d'une part avec le Modèle 1, pour $n = N = 100$, $\widehat{\text{Err}}(\hat{\sigma}^2_{\texttt{MS}}) = 0.03$ et pour $n = N = 1000$, $\widehat{\text{Err}}(\hat{\sigma}^2_{\texttt{MS}}) = 0.006$. D'autre part, nous avons avec le Modèle 2 pour $n = N = 100$, $\widehat{\text{Err}}(\hat{\sigma}^2_{\texttt{MS}}) = 0.61$ et pour $n = N = 1000$, $\widehat{\text{Err}}(\hat{\sigma}^2_{\texttt{MS}}) = 0.43$. Finalement, nous concluons que plus la fonction de variance prend de grandes valeurs, plus son estimation devient difficile.

Bibliographie

Anderson, T.G. et Lund, J. (1997). Estimating continuous-time stochastic volatility models of the short- term interest rate, *Journal of Econometrics*, 77(2) :343–377.

Denis, C., Hebiri, M. et Zaoui, A. (2020). Regression with reject option and application to knn. *NeurIPS 2020*.

Fan, J., et Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3) :645–660.

Hart, J. (1997). Nonparametric Smoothing and Lack-of-Fit Tests. *Springer Series in Statistics*.

Härdle, W., et Tsybakov, A.B. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81(1) :223–242.

Kulik, R., et Wichelhaus, C. (2011). Nonparametric conditional variance and error density estimation in regression models with dependent errors and predictors. *Electron. J. Statist.*, 5 :856–898.

Nemirovski, A. (2000). Topics in Non-parametric Statistics. *Saint-Flour Summer School in Probability XXVIII, 1998. Lecture Notes in Mathematics 1738. Springer, NY.*

Ruppert, D., Wand, M.P., Holst, U. et HöSJER, O.(1997). Local polynomial variance function estimation. *Technometrics*, 39(3) :262–273.

Tsybakov, A.B. (2003). Optimal rates of aggregation. *Learning Theory and Kernel Machines*, 303–313.

Tsybakov, A.B. (2014). Aggregation and minimax optimality in high-dimensional estimation. *Proceedings of International Congress of Mathematicians*, 3 :225–246.

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, 10 :25–47.

Construction d'histogrammes irréguliers selon le principe MDL

Valentina ZELAYA MENDIZABAL [1] & Marc BOULLE [2] & Fabrice ROSSI [3]

[1] *Orange Labs, Université Panthéon-Sorbonne,*
valentina.zelayamendizabal@orange.com
[2] *Orange Labs,* *marc.boulle@orange.com*
[3] *Université Paris-Dauphine,* *Fabrice.Rossi@dauphine.psl.eu*

Résumé. Nous proposons dans cette communication une méthode de construction totalement automatique d'histogrammes irréguliers basée sur le principe du *Minimum Description Length*. Couplée à une heuristique d'optimisation, notre approche permet une construction en $\mathcal{O}(n \log n)$ pour n observations ce qui la rend applicable à des données très volumineuses, contrairement aux méthodes existantes de même nature. Une évaluation expérimentale sur des données synthétiques et réelles montre les atouts de notre approche par rapport à celles de l'état de l'art.

Mots-clés. Histogramme, estimation de densité, sélection de modèle

Abstract. We present in this paper a new fully automated method for irregular histogram construction based on the *Minimum Description Length* principle. Associated to a greedy search heuristic, our method scales in $\mathcal{O}(n \log n)$ for n observations and can be applied to large scale data sets, contrarily to existing work. An experimental evaluation on synthetic and real data shows the strengths and limitations of our approach compared to state-of-the-art methods.

Keywords. Histograms, density estimation, model selection

1 Construction automatique d'histogrammes

Malgré leur défauts, les histogrammes restent un outil populaire d'estimation de densité, notamment en raison de leur interprétation visuelle simple. Leur emploi systématique est notamment facilité par l'utilisation de règles simples pour les construire : beaucoup de logiciels statistiques se contentent de proposer des histogrammes réguliers, avec des intervalles de longueurs égales, et de choisir le nombre d'intervalles au moyen de règles simples comme la vénérable règle de Sturges [8] ou celle de Freedman-Diaconis [3]. Cependant pour des distributions complexes, comme celles à queues lourdes, des histogrammes à intervalles de longueurs variables sont plus adaptés et il existe assez peu de méthodes complètement automatiques pour adapter le nombre d'intervalles et les points de coupure aux données (cf [2, 6] pour deux états de l'art sur ces méthodes).

Nous proposons dans cette communication une nouvelle méthode de ce type. Comme dans la plupart des travaux s'attaquant à ce problème, nous considérons la construction

d'un histogramme comme un problème de sélection de modèle. Ce type de problème est généralement résolu par une stratégie de maximum de vraisemblance pénalisé. Les pénalités retenues dans les méthodes les plus efficaces comme [6] s'appuient sur des résultats théoriques mais aussi sur des considérations heuristiques et expérimentales.

Nous utilisons dans cette communication une approche similaire basée sur le principe du *Minimum Description Length* (MDL) et notamment sur le « maximum de vraisemblance normalisé » utilisé dans [4] pour dériver un critère de sélection de modèle (*Normalized Maximum Likelihood*, `NML`) adapté aux histogrammes irréguliers. Ce critère donne des résultats convaincants sur des distributions simples mais présente deux limitations : il dépend d'un paramètre de précision et son optimisation est coûteuse en temps de calcul.

Nous proposons un nouveau critère inspiré de `NML` et de type MDL, qui l'améliore sur deux points. Dans sa version simple, nommée `Enum`, ce critère est plus rapide à calculer que le `NML`. Dans sa version plus complexe, nommée `G-Enum`, le critère permet d'automatiser le choix de la précision de l'estimation. Nous proposons en outre une heuristique d'optimisation qui conduit à une construction d'un histogramme en $\mathcal{O}(n \log n)$ pour n observations (contre un temps polynomial pour `NML`).

Nous consacrons la suite de cette communication à la présentation du critère proposé et à celle d'une évaluation comparative de ses performances comparées à l'état de l'art.

2 Histogrammmes G-Enum

2.1 Formulation du problème

Nous considérons un échantillon de n observations $x^n = (x_1, ..., x_n)$ sur l'intervalle $[x_{\min}, x_{\max}]$. On note ϵ la précision de représentation des données : il s'agit du paramètre à régler dans le `NML`. Chaque observation $x_j \in x^n$ est approchée par un élément de $\widetilde{x}_j \in \mathcal{X} = \{x_{\min} + t\epsilon; t = 0, ..., E\}$ où $E = \dfrac{x_{\max} - x_{\min}}{\epsilon}$.

On considère des histogrammes construits à partir de la grille $\mathcal{C}$ obtenue à partir des milieux de paires de valeurs consécutives de $\mathcal{X}$, comme dans [4]

$$\mathcal{C} = \{x_{\min} + \epsilon/2 + t\epsilon; t = 0, ..., E-1\},$$

avec $c_0 = x_{\min} - \epsilon/2$ et $c_K = x_{\max} + \epsilon/2$. Ces points de coupure définissent E *cellules élémentaires* de longueur ϵ, que nous appelons ϵ-bins (voir figure 1). Ils constituent les éléments de base des intervalles d'un histogramme : chaque combinaison de ϵ-bins en K intervalles, avec K allant de 1 à E, définit un modèle d'histogramme. Dans cet éventail de possibilités, notre objectif est de sélectionner un ensemble de $K-1$ points de coupure $C = (c_1, ..., c_{K-1})$, $c_k \in \mathcal{C}$ tels que $[x_{\min} - \epsilon/2, x_{\max} + \epsilon/2]$ soit partitionné en K intervalles $\{[c_0, c_1],]c_1, c_2], ...,]c_{K-1}, c_K]\}$ adaptés à la distribution réelle des données (voir figure 1). Chaque intervalle k a un effectif de h_k observations et une longueur $L_k = c_k - c_{k-1}$, qui est un multiple de ϵ : $\forall k,\ \exists\ E_k \in \mathbb{N}^*$ tel que $L_k = E_k \cdot \epsilon$.

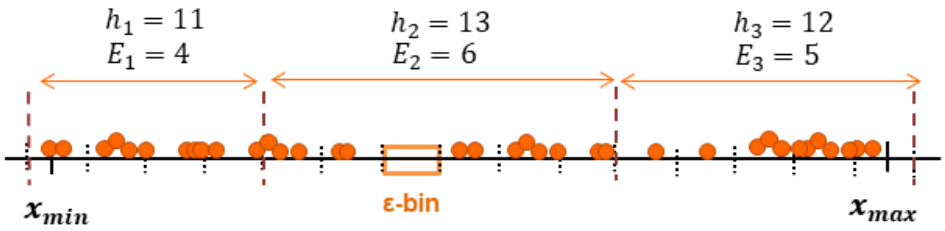


FIGURE 1 – Un choix possible d'intervalles, avec leurs effectifs et longueurs

Un histogramme est entièrement défini par le choix du nombre d'intervalles, l'ensemble des points de coupure qui les définissent et les effectifs associés.

2.2 Formalisation d'un critère granularisé pour les histogrammes

Les critères proposés sont donnés sans justification dans la table 1 pour des raisons de place. Ils sont obtenus en interprétant l'approche MDL comme une recherche de maximum a posteriori. Chaque critère comprend un terme de vraisemblance et un terme de codage ou d'a priori sur les paramètres du modèle (ici l'histogramme). Dans la table, ces deux termes sont réorganisés pour faire apparaître les différences entre les trois critères.

Le critère `Enum` est directement issu de ce point de vue bayésien sur le MDL et préserve certaines propriétés d'optimalité du Maximum de vraisemblance Normalisé [1]. Par rapport au critère `NML`, il évite un terme complexe à calculer, log $\mathcal{R}^n_{\mathcal{M}}$ la complexité paramétrique [4]. En outre la pénalisation induite est croissante avec le nombre d'intervalles.

Afin d'automatiser le choix du seul paramètre des méthodes `NML` et `Enum`, nous introduisons une version avec une granularité G à optimiser : un intervalle d'histogramme est maintenant constitué de G_k g-bins, chaque g-bins étant elle même constituée de $g = \frac{E}{G}$ ϵ-bins. On découple ainsi la précision de représentation des données (les ϵ-bins) de la précision de représentation des histogrammes (les g-bins). On peut ainsi fixer ϵ autour de la précision machine, g étant optimisé dans la procédure de sélection de modèle. Le critère obtenu est baptisé `G-Enum`.

2.3 Heuristiques de recherche

L'algorithme de programmation dynamique proposé pour optimiser le critère `NML` est optimal, mais il nécessite un temps de calcul en $\mathcal{O}(E^3)$ [4]. Nous utilisons diverses heuristiques de recherche, s'appuyant notamment sur l'additivité des critères et sur une stratégie gloutonne de fusion d'intervalles adjacents. La complexité est ainsi réduite en $\mathcal{O}(n \log n)$. Pour limiter l'impact de cette approche sur la qualité des histogrammes obtenus, nous utilisons des post-optimisations heuristiques comme des découpages et combinaisons des intervalles une fois un optimum local atteint.

TABLE 1 – Comparaison des termes entre critères

Critère	Termes d'indexation	Termes multinomiaux	Termes d'indexation des bins
NML [4]	$\log \binom{E}{K-1}$	$\log\ \mathcal{R}^n_{\mathcal{M}} + \log \frac{n^n}{{h_1}^{h_1} \cdots {h_K}^{h_K}}$	$\sum_{k=1}^{K} h_k \log E_k$
Enum	$\log^* K + \log \binom{E+K-1}{K-1}$	$\log \binom{n+K-1}{K-1} + \log \frac{n!}{h_1!...h_K!}$	$\sum_{k=1}^{K} h_k \log E_k$
G-Enum	$\log^* K + \log^* G + \log \binom{G+K-1}{K-1}$	$\log \binom{n+K-1}{K-1} + \log \frac{n!}{h_1!...h_K!}$	$\sum_{k=1}^{K} h_k \log G_k + n \log \frac{E}{G}$

où $\log^* K$ est le codage universel des nombres entiers proposé par Rissannen [5]

3 Évaluation expérimentale

3.1 Protocole et métriques

Nous évaluons notre stratégie par comparaison avec une sélection de méthodes concurrentes automatiques (règles de Sturges et Freedman-Diaconis [3], taut strings [2], RMG [6] et Bayesian blocks [7]) sur plusieurs échantillons de 6 types de distribution. Ces méthodes sont testées sur des échantillons de tailles croissantes, allant de $n = 10$ à $n = 10^5$ ou $n = 10^6$.

Les résultats sont comparés selon trois métriques : le nombre d'intervalles, le temps de calcul et la distance de Hellinger par rapport au modèle de distribution. Afin d'évaluer la variabilité des résultats, nous présentons les moyennes et les écarts types calculés sur un total de 10 d'expériences pour une distribution donnée et une taille d'échantillon donnée.

On donne ici un extrait de cet ensemble volumineux d'expériences (les tests sur des données réelles allant jusqu'à 25 millions d'observations ne sont pas présentés).

3.2 Résultats clés

Comparaison entre méthodes MDL

Les histogrammes `NML` et `Enum` sont interchangeables en termes de nombre d'intervalles et de distance de Hellinger, et ce quel que soit l'algorithme choisi pour optimiser les critères. Toutefois, en termes de temps de calcul, il y a un avantage significatif à préférer

l'heuristique de recherche et les critères énumératifs plus simples. Les histogrammes `G-Enum` prennent un peu plus de temps à calculer mais produisent des estimations légèrement meilleures en termes de distance de Hellinger et ne nécessitent pas de fixer ϵ de façon spécifique.

Comparaison avec d'autres méthodes de l'état de l'art (tables 2 et 3)

Les autres méthodes ont des meilleurs résultats dans les cas spécifiques pour lesquels elles ont été conçues. Bien qu'ils soient rarement en première place pour chaque type de distribution, les histogrammes `G-Enum` sont toujours parmi les meilleurs estimateurs, et ce sans la forte variabilité des autres méthodes. Parmi les histogrammes irréguliers, `G-Enum` se distingue comme le plus parcimonieux en nombre d'intervalles. Il s'agit d'une qualité importante pour l'analyse exploratoire car cela facilite l'interprétation des résultats. `G-Enum` est également de loin la plus rapide des méthodes irrégulières, ce qui la rend adaptée aux jeux de données de grande taille.

Références

[1] M. Boullé, F. Clérot, and C. Hue. Revisiting enumerative two-part crude MDL for Bernoulli and multinomial distributions (extended version). Technical report, arXiv, abs/1608.05522, 2016.

[2] Davies, Laurie, Gather, Ursula, Nordman, Dan, and Weinert, Henrike. A comparison of automatic histogram constructions. *ESAIM : PS*, 13 :181–196, 2009.

[3] David Freedman and Persi Diaconis. On the histogram as a density estimator :l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4) :453–476, Dec 1981.

[4] Petri Kontkanen and Petri Myllymäki. MDL histogram density estimation. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 219–226, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[5] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, 11(2) :416–431, 06 1983.

[6] Yves Rozenholc, Thoralf Mildenberger, and Ursula Gather. Combining regular and irregular histograms by penalized likelihood. *Computational Statistics and Data Analysis*, 54(12) :3313 – 3323, 2010.

[7] Jeffrey D. Scargle, Jay P. Norris, Brad Jackson, and James Chiang. Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations. *Astrophysical Journal*, 764(2) :167, February 2013.

[8] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153) :65–66, 1926.

TABLE 2 – Comparaison des distances de Hellinger sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [7]	TS [2]	RMG [6]	FD [3]	Sturges
Normale	$0.045 \pm 6 \cdot 10^{-4}$	0.046 ± 0.002	0.047 ± 0.002	0.040 ± 0.002	0.034 ± 0.002	0.033 ± 0.002	0.055 ± 0.002
Cauchy	0.061 ± 0.004	0.074 ± 0.003	0.064 ± 0.002	0.045 ± 0.005	0.064 ± 0.001	0.138 ± 0.002	0.862 ± 0.036
Uniforme	0.024 ± 0.001	0.050 ± 0.005	0.025 ± 0.004	0.031 ± 0.015	0.029 ± 0.011	0.082 ± 0.011	0.028 ± 0.002
Triangle	0.039 ± 0.002	0.038 ± 0.0025	0.039 ± 0.001	0.084 ± 0.024	0.084 ± 0.029	0.032 ± 0.002	$0.049 \pm 9 \cdot 10^{-4}$
4 triangles	0.037 ± 0.002	0.038 ± 0.003	0.040 ± 0.003	0.078 ± 0.029	0.069 ± 0.026	0.032 ± 0.002	$0.043 \pm 4 \cdot 10^{-4}$
6 gaussiennes	0.057 ± 0.002	0.059 ± 0.002	0.060 ± 0.002	0.040 ± 0.001	0.052 ± 0.002	0.060 ± 0.001	0.142 ± 0.013

TABLE 3 – Comparaison des temps de calcul (en secondes) sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [7]	TS [2]	RMG [6]	FD [3]	Sturges
Normale	0.014 ± 0.003	2.724 ± 0.283	5.785 ± 0.479	0.014 ± 0.002	1.239 ± 0.085	0.002 ± 2.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
Cauchy	0.028 ± 0.006	121.60 ± 4.99	3.250 ± 0.112	0.116 ± 0.205	0.906 ± 0.107	0.009 ± 0.014	0.001 ± 0.003
Uniforme	0.015 ± 0.002	0.168 ± 0.012	5.989 ± 0.167	0.011 ± 0.002	1.387 ± 0.139	0.002 ± 3.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
Triangle	0.014 ± 0.005	0.169 ± 0.005	5.962 ± 0.113	0.015 ± 0.002	1.291 ± 0.091	0.002 ± 2.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
4 triangles	0.012 ± 0.006	0.103 ± 0.027	3.004 ± 0.245	0.013 ± 0.006	0.954 ± 0.138	0.002 ± 0.005	0.0 ± 0.0
6 gaussiennes	0.017 ± 0.002	1.91 ± 0.085	4.165 ± 0.369	0.048 ± 0.005	1.056 ± 0.077	0.006 ± 0.008	0.002 ± 0.005

TABLE 4 – Comparaison du nombre d'intervalles sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [7]	TS [2]	RMG [6]	FD [3]	Sturges
Normale	16.30 ± 0.46	15.60 ± 1.02	15.90 ± 1.04	72.50 ± 4.41	39.70 ± 7.34	62.40 ± 2.29	15.0 ± 0.0
Cauchy	30.90 ± 2.43	23.60 ± 1.02	29.40 ± 1.56	144.90 ± 9.26	29.50 ± 1.57	110711.90 ± 132580.43	15.0 ± 0.0
Uniforme	1.0 ± 0.0	2.80 ± 0.60	1.30 ± 0.90	3.70 ± 5.44	1.70 ± 1.80	22.0 ± 0.0	15.0 ± 0.0
Triangle	12.50 ± 0.92	13.60 ± 0.66	12.60 ± 0.66	48.0 ± 5.85	33.70 ± 8.25	32.10 ± 0.30	15.0 ± 0.0
4 triangles	11.20 ± 0.75	12.00 ± 0.77	10.90 ± 0.70	42.30 ± 4.90	27.60 ± 8.0	30.80 ± 0.40	15.0 ± 0.0
6 gaussiennes	28.90 ± 1.22	27.30 ± 1.49	27.00 ± 2.00	134.90 ± 9.37	100.40 ± 11.60	66.40 ± 2.42	15.0 ± 0.0

Une nouvelle dissimilarité pour le partitionnement spatial de pluies extrêmes, non-paramétrique et liant théorie des valeurs extrêmes bivariée et marginales

Margaux Zaffran [1,2,◇] & Philippe Naveau [3]

[1] *INRIA, Sophia-Antipolis*
[2] *EDF R&D, 7 bd Gaspard Monge, 91120 Palaiseau, France*
[◇] *travail réalisé durant un stage au Laboratoire des Sciences du Climat et de l'Environnement*
margaux.zaffran@inria.fr
[3] *Laboratoire des Sciences du Climat et de l'Environnement, LSCE-IPSL-CNRS, Gif-sur-Yvette, France.*
philippe.naveau@lsce.ipsl.fr

Résumé. Les évènements climatiques extrêmes ont un impact sociétal important, particulièrement les fortes précipitations. De récents événements dramatiques, comme la tempête Alex qui a frappé le nord de l'Italie et le sud-est de la France, mettent en lumière l'intérêt de mieux connaitre ce type d'évènement extrêmes afin d'adapter les infrastructures. Au regard du sous-échantillonage inhérent aux extrêmes, la création de régions cohérentes permettrait d'améliorer l'inférence des modélisations de ces extrêmes. Nous proposons une nouvelle dissimilarité adaptée aux extrêmes, fondée sur les distributions de probabilités bivariées et des marginales, que nous appliquons à un algorithme de partitionnement. Les résultats obtenus sur des données quotidiennes françaises, de 1976 à 2015, sont cohérents d'un point de vue climatologique. L'algorithme développé est disponible sous forme d'un package R.

Mots-clés. partitionnement, valeurs extrêmes, Théorie des Valeurs Extrêmes, divergence de Kullback-Leibler, dépendance de queue, précipitations, climat

Abstract. Climate extremes have a strong societal impact, especially extreme rainfalls. Recent dramatic events, like Alex Storm that stroke northern Italy and South-East of France, emphasize the need of a good knowledge of extreme events to adapt infrastructures. In light of subsampling inherent to extremes, the creation of coherent regions would improve the inference of models for these extremes. We propose a new dissimilarity tailored for extremes, based on bivariate and marginals probability distributions, and plug it into a clustering algorithm. Obtained results on daily French data, from 1976 to 2015, are climatologically consistent. The developed algorithm is available in the form of an R package.

Keywords. clustering, extreme values, Kullback-Leibler divergence, tail dependence, rainfall, climate

1 Introduction

De récents événements dramatiques, comme la tempête Alex qui a frappé le nord de l'Italie et le sud-est de la France durant l'automne 2020, mettent en lumière l'intérêt de mieux modéliser les évènements extrêmes de pluie afin d'adapter les infrastructures. Le cadre de la théorie des valeurs extrêmes (EVT, Fisher et Tippett (1928), Pickands (1975), Coles (2001) et plus récemment Davison et Huser (2015)) est fréquemment utilisé pour ces applications climatiques. La question est généralement : quelle est la distribution de probabilité sous-jacente aux pluies extrêmes dans un lieu donné (par exemple, Marseille) ? Le peu de données disponibles pour les extrêmes amène naturellement à chercher des groupes de stations météo semblables dans leurs extrêmes, afin de les regrouper et améliorer l'inférence de leurs modélisations.

Nous nous intéressons dans cette étude au développement d'une telle méthode de partitionnement, par la construction d'une dissimilarité adaptée. La définition de stations similaires peut prendre principalement deux directions : la dépendance temporelle (une forte pluie à Marseille s'accompagne généralement d'une forte pluie à Aix-En-Provence) ou la loi de distribution marginale de ces extrêmes. Les approches proposées dans Bernard et al. (2013), Bador et al. (2015), Saunders et al. (2020) s'intéressent à la dépendance temporelle. Les études axées "Regional Frequency Analysis" (RFA), comme Carreau et al. (2017), s'intéressent, elles, plutôt uniquement aux distributions des lois marginales. Ici, nous proposons une méthode combinant les deux concepts.

Notre objectif est de développer une telle méthode qui soit non-paramétrique et qui puisse passer à l'échelle convenablement sur de grands jeux de données. Nous avons à notre disposition un jeu de données Météo-France contenant les volumes de précipitations quotidiennes dans 174 stations, de 1976 jusqu'à 2015.

FIGURE 1 – Couverture des 174 stations Météo-France stations constituant le jeu de données de précipitations quotidiennes, du 01/01/1976 au 31/12/2015. Les différentes couleurs représentent les grandes régions climatiques de pluies extrêmes, d'après *Pluies extrêmes en France métropolitaine : un peu de géographie* (p. d.).

Sur la figure 1 nous superposons les principaux climats de pluies extrêmes françaises. Ce découpage a été obtenu par la connaissance climatique du terrain. Cependant, en pratique, nous n'avons pas toujours cette connaissance (une maille plus fine, des résultats de simulations de précipitations futures ou tout simplement nous n'y avons pas facilement accès). De plus, ce découpage est en réalité plus flou que sur notre figure 1 : les zones adjacentes à plusieurs régions ne sont pas classifiables. Ainsi, notre objectif est justement de proposer une méthode non-supervisée qui renvoie un tel découpage. Le fait de l'appliquer à ces données nous permettra de nous comparer au découpage de Météo-France.

2 Méthode

Nous proposons une dissimilarité prenant en compte d'une part la proximité des lois marginales des excès (proche de ce qui est fait en RFA), et d'autre part la dépendance temporelle de ces excès (similaire à la dissimilarité utilisée dans Bernard et al. (2013), Bador et al. (2015), Saunders et al. (2020)).

La proximité des lois marginales est évaluée à l'aide de l'estimateur de la Kullback-Leibler des excès, proposé par Naveau et al. (2014). Cet estimateur estime la quantité définie en (1), qui compare les variables aléatoires X et Y, de fonction de répartition respectives F et G, au-dessus d'un seuil u.

$$K(f_u, g_u) = -L(f_u; g_u) - L(g_u; f_u) \tag{1}$$

avec :

$$L(f_u; g_u) = \mathbb{E}_f \left[\log \left(\frac{\overline{G}(X)}{\overline{G}(u)} \right) | X > u \right] + 1$$

La dépendance temporelle est quant à elle estimée via le coefficient de dépendance de queue résiduel, $\bar{\chi}$, introduit par Ledford et Tawn (1996) (voir également Coles et al. (1999)). Pour X et Y deux variables aléatoires de fonction de répartition respectives F et G, nous estimons $\bar{\chi}(q)$ défini dans (2).

$$\bar{\chi} = \lim_{q \to 1} \bar{\chi}(q) = \lim_{q \to 1} \frac{\log\left(\mathbb{P}\{X > F^{-1}(q)\}\mathbb{P}\{Y > G^{-1}(q)\}\right)}{\log\left(\mathbb{P}\{X > F^{-1}(q), Y > G^{-1}(q)\}\right)} - 1 \tag{2}$$

Deux types de paramètres sont alors à choisir : les seuils définissant les excès (u et q) pour obtenir des estimateurs, et le paramètre λ représentant le poids de chacune de ces deux comparaisons dans la dissimilarité finale.

Pour des raisons d'interprétabilité, de praticité du choix des paramètres et de passage à l'échelle, nous utilisons l'algorithme des k-médoides (ou PAM, Partitioning Around Medoids), proposé par Kaufman et Rousseeuw (1987).

La figure 2 synthétise le déroulement de notre procédure, les éléments verts représentant l'apport proposé.

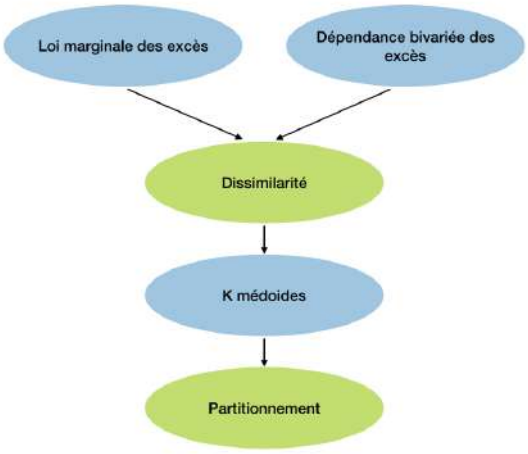


FIGURE 2 – Schéma de la méthode proposée. Les éléments verts indiquent la contribution.

Nous proposons enfin une procédure pour choisir les paramètres de seuil et λ, basée sur l'indice de Rand (voir Rand (1971)).

Les détails techniques de cet exposé seront bientôt disponibles dans Zaffran et Naveau (2021).

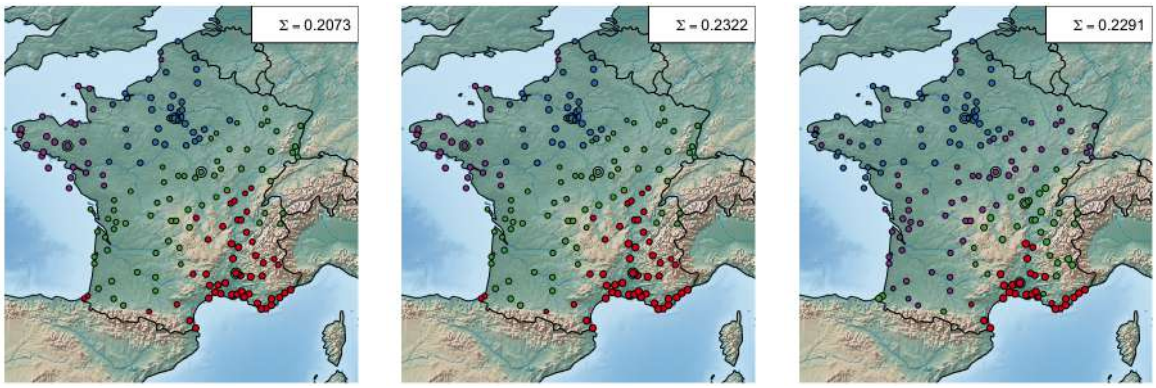
3 Résultats

Les résultats obtenus, dont un échantillon est représenté en figure 3, sont cohérents avec les climats français : nous retrouvons nettement la vallée du Rhône et le bassin méditerranéen (groupe rouge), le Nord et le bassin parisien (groupe bleu). L'Aquitaine est conservée au sein d'un même groupe en permanence (groupe vert ou violet).

Le paramètre λ représente le poids de la contribution de la dépendance des marginales : plus il est élevé, plus elle est importante et moins la contribution de la dépendance temporelle l'est.

Lorsque nous comparons avec les résultats obtenus via la méthode de Bernard et al. (2013), nous remarquons que nous obtenons des groupes beaucoup moins compacts spatialement. La prise en compte des lois marginales permet de retrouver plus finement les régions climatiques que Bernard et al. (2013), par rapport à la carte de Météo-France, voir figure 1. En effet, nous retrouvons le climat méditerranéen (groupe rouge) ainsi que le climat océanique plus ou moins altéré (groupe vert puis violet).

D'autre part, l'intégration de la dépendance temporelle amène plus de stabilité et de compacité aux bords : pour $\lambda = 0.75$ (figure 3c) le groupe violet semble moins cohérent climatiquement que les groupes violets et verts pour des λ plus faibles (figures 3b et 3a),



(a) $\lambda = 0.25$ (b) $\lambda = 0.5$ (c) $\lambda = 0.75$

FIGURE 3 – Partitionnements obtenus pour différentes valeurs de λ dans la dissimilarité. Σ en légende correspond au coefficient de silhouette (Rousseeuw, 1987) moyen : plus il est proche de 1, mieux sont classifiés les points, et inversement si il est proche de -1.

c'est-à-dire lorsqu'on augmente la contribution de la dépendance temporelle.

L'analyse de la stabilité des groupes obtenus lorsqu'on augmente le seuil u, basée sur une quantité proche de l'indice de Rand (Rand, 1971), nous fait préférer le partitionnement de la figure 3b.

4 Conclusion

Nous avons proposé une méthode non-paramétrique et adaptable pour des grands jeux de données qui retrouve avec succès des groupes climatologiquement cohérents. Cette méthode mélange un critère d'extrêmes bivariés (dépendance temporelle) à un critère sur les lois marginales. Notre algorithme a été implémenté dans un package `R`.

La méthode pourrait être facilement appliquée à d'autres jeux de données, comme :

— des résultats de simulation climatiques pour analyser l'évolution des groupes avec le temps ;

— des données plus locales, comme la région de Montpellier et des Cévennes, qui est le berceau d'épisodes de précipitations très intenses, appelées les "pluies cévenoles" ;

— d'autres types de données, comme le vent ou la température, grâce aux faibles hypothèses de notre procédure.

Bibliographie

Bador, M. et al. (2015). "Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe". In : *Weather and Cli-*

mate Extremes 9. The World Climate Research Program Grand Challenge on Extremes & WCRP-ICTP Summer School on Attribution and Prediction of Extreme Events, p. 17-24. ISSN : 2212-0947. DOI : https://doi.org/10.1016/j.wace.2015.05.003.

Bernard, E. et al. (2013). "Clustering of Maxima : Spatial Dependencies among Heavy Rainfall in France". In : *Journal of Climate* 26.20, p. 7929-7937. ISSN : 0894-8755. DOI : 10.1175/JCLI-D-12-00836.1.

Carreau, J. et al. (2017). "Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation". In : *Water Resources Research* 53. DOI : 10.1002/2017WR020758.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. T. 208. Springer.

Coles, S. et al. (1999). "Dependence measures for extreme value analyses". In : *Extremes* 2.4, p. 339-365.

Davison, A. C. et R. Huser (2015). "Statistics of extremes". In : *Annual Review of Statistics and its Application* 2, p. 203-235.

Fisher, R. A. et L. H. C. Tippett (1928). "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In : *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2, p. 180-190. DOI : 10.1017/S0305004100015681.

Kaufman, L. et P. J. Rousseeuw (1987). "Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm". In : *Y. Dodge, Ed*, p. 405-416.

Ledford, A. W. et J. A. Tawn (1996). "Statistics for near independence in multivariate extreme values". In : *Biometrika* 83.1, p. 169-187.

Naveau, P. et al. (2014). "A non-parametric entropy-based approach to detect changes in climate extremes". In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 76.5, p. 861-884. DOI : 10.1111/rssb.12058.

Pickands, J. (1975). "Statistical Inference Using Extreme Order Statistics". In : *Ann. Statist.* 3.1, p. 119-131. DOI : 10.1214/aos/1176343003.

Pluies extrêmes en France métropolitaine : un peu de géographie (p. d.), last accessed February 21, 2021. URL : http://pluiesextremes.meteo.fr/france-metropole/Un-peu-de-geographie.html.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In : *Journal of the American Statistical Association* 66.336, p. 846-850. DOI : 10.1080/01621459.1971.10482356.

Rousseeuw, P. J. (1987). "Silhouettes : A graphical aid to the interpretation and validation of cluster analysis". In : *Journal of Computational and Applied Mathematics* 20, p. 53-65. ISSN : 0377-0427. DOI : https://doi.org/10.1016/0377-0427(87)90125-7.

Saunders, K. et al. (2020). "A regionalisation approach for rainfall based on extremal dependence". In : *Extremes*. DOI : 10.1007/s10687-020-00395-y. URL : https://doi.org/10.1007/s10687-020-00395-y.

Zaffran, M. et P. Naveau (2021). "Spatial clustering of rainfall extremes by coupling Kullback-Leibler divergence and tail coefficient". In : *rapport interne du LSCE (à soumettre prochainement)*.

Modèle des queues proportionnelles pour l'estimation de quantiles extrêmes.

Benjamin Bobbia [1] & Clément Dombry [2] & Davit Varron [3]

[1] *Laboratoire de mathématiques de Besançon, 16 route de gray, Besançon.*
benjamin.bobbia@univ-fcomte.fr
[2] *Laboratoire de mathématiques de Besançon, 16 route de gray, Besançon.*
clement.dombry@univ-fcomte.fr
[3] *Laboratoire de mathématiques de Besançon, 16 route de gray, Besançon.*
davit.varron@univ-fcomte.fr

Résumé. Nous présentons ici un modèle qui vise l'estimation de quantiles extrêmes dans le cadre de la régression. Plus précisément, nous nous concentrons sur l'estimation de quantiles élevés de la distribution d'une variable Y en fonction d'une configuration de covariable donnée dans $\mathbb{R}^d$. Le modèle présenté s'inspire de celui des extrêmes hétéroscedastiques. Il consiste à supposer la proportionalité asymptotique entre la queue de distribution de Y et celle de Y sachant $X = x$. Nous construisons et étudions la normalité asymptotique d'estimateurs des paramètres du modèle pour en déduire les propriétés asymptotiques d'un estimateur du quantile de type Weissman.

Mots-clés. valeurs extrêmes, estimation quantile, mesures empiriques.

Abstract. We are interested in extreme quantile estimation in the regression framework. We estimate high quantile of the distribution of a random variable Y given a certain setup of covariate X in $\mathbb{R}^d$. The model presented here is inspired by the heteroscedastic extremes and consist to assume the asymptotic proportionality between the tail distribution of Y and the conditional tail distribution of Y given $X = x$. We build estimators of each model parameter and study the asymptotic normality. Consequently we deduce asymptotic properties about a Weissman-Type quantile estimate.

Keywords. Extreme values, Quantile estimation, Empirical measure.

1 Position du problème

L'objectif de ce travail est l'estimation des quantiles d'une variable aléatoire réelle Y conditionnellement aux valeurs prises par un vecteur aléatoire X dans $\mathbb{R}^d$. En notant F_x la fonction de répartition de Y sachant $X = x \in \mathbb{R}^d$, le quantile conditionnel d'ordre α_n est défini par

$$q(\alpha|x) = F_x^{\leftarrow}(1 - \alpha_n),$$

avec $F_x^{\leftarrow}$ l'inverse généralisée de F_x. Dans cette étude, nous nous intéressons au cas où α_n tend vers 0 lorsque n croît. De tels quantiles sont qualifiés d'extrêmes. Dans ce contexte il est alors nécessaire d'extrapoler les queues de distributions (conditionnelles et non conditionnelles) de Y.

En l'absence de covariables le problème est déjà bien étudié. Par exemple, I Weissman utilise en 1978 l'approximation Pareto des distributions à queues lourdes pour estimer de tels quantiles. Il existe d'autres exemples d'estimateurs des quantiles extrêmes, que l'on peut retrouver dans la monographie de de Haan et Ferreira par exemple. Ce cas étant connu, l'objectif pour estimer des quantiles conditionnels est alors de proposer un modèle dans lequel il est possible de relier les queues de distribution conditionnelle et non conditionnelle. Le modèle que nous proposons alors est inspiré du modèle des extrêmes hétéroscedastiques proposé par John Einmahl, Laurens de Haan et Chen Zhou de 2016 au sens où nous supposons, asymptotiquement, l'existance d'un coefficient de proportionnalité entre la queue de distribution de Y conditionnellement à $X = x$ et la distribution de Y non conditionnelle.

2 Le modèle des queues proportionnelles

2.1 Présentation du modèle

Soient (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathbb{R}$. L'hypothèse principale du *modèle des queues proportionnelles* est

$$\lim_{y \to +\infty} \frac{1 - F_x(y)}{1 - F(y)} = \sigma(x) \quad \text{uniformément en } x \in \mathbb{R}^d, \tag{1}$$

oú σ est appelé fonction skedasis et F désigne la fonction de répartition de Y. Par intégration, on remarque que σ est une $\mathbf{P}_X$-densité.

La seconde hypothèse est la bien nommée condition des valeurs extrêmes du premier ordre. Comme nous nous concentrons sur le cas des distributions à queues lourdes, cette condition peut alors être exprimée comme suit :

$$\lim_{t \to \infty} \frac{U(tx)}{U(t)} = x^{-1/\gamma} \quad , x \geq 1,$$

avec $U(t) = F^{\leftarrow}(1 - 1/t)$ et $\gamma > 0$.
Ces deux hypothèses combinées conduisent au fait que toutes les distributions conditionnelles F_x sont à queues lourdes avec le même indice des valeurs extrêmes γ. Autrement dit, γ ne dépend pas de $x \in \mathbb{R}^d$. Ce qui est peu restrictif dans le cas des distributions à queues lourdes. En effet, dans la pratique, on pourrait ajuster le modèle des queues proportionnelles localement, sur des zones où l'indice des valeurs extrêmes varie peu et peut

alors être assimilé à une constante. De plus, cette conséquence nous fournit un moyen commode de tester l'adéquation d'un jeu de données à ce modèle. Il est difficile de vérifier si un jeu de données satisfait l'hypothèse de proportionnalité des queues, alors qu'il est plus simple de vérifier si l'indice γ reste constant sur $\mathbb{R}^d$.

Dans ce modèle, le comportement des extrêmes est dirigé par deux paramètres. L'indice des valeurs extrêmes $\gamma > 0$, duquel dépend leur intensité, et σ qui influence leur répartition ainsi que leur fréquence. Il est donc naturel d'estimer ces deux paramètres. De plus, nous considérerons également la version intégrée de la fonction skedasis,

$$C(x) = \int_{\{u \leq x\}} \sigma(u) \mathbf{P}_X(du), \quad x \in \mathbb{R}^d, \tag{2}$$

avec $u \leq x$ la comparaison de vecteurs composante par composante. Cette fonction, bien plus facile à estimer que σ, nous permettra de construire des tests alors que la fonction skedasis interviendra dans l'estimation de quantile.

Pour des raisons techniques, il est nécessaire de faire des hypothèses sur les vitesses de convergence dans nos hypothèses. Plus exactement, supposons l'existence d'une fonction A, tendant vers 0 lorsque $y \to \infty$, telle que

$$\sup_{x \in \mathbb{R}^d} \left| \frac{1 - F_x(y)}{\sigma(x)(1 - F(y))} - 1 \right| = \mathrm{O}\left(A\left(\frac{1}{1 - F(y)} \right) \right) \tag{3}$$

et

$$\sup_{z > \frac{1}{2}} \left| \frac{1 - F(zy)}{z^{-\alpha}(1 - F(y))} - 1 \right| = \mathrm{O}\left(A\left(\frac{1}{1 - F(y)} \right) \right). \tag{4}$$

Remarquons que les hypothèses (1) et (3) sont en réalité équivalentes. Seule l'hypothèse (4) est restrictive. En effet, cette hypothèse est plus forte que les classiques conditions du premier ordre en extrême mais tout de même plus souple que les conditions du second ordre.

2.2 Estimation des paramètres

Soient $(X_i, Y_i)_{1 \leq i \leq n}$ des copies i.i.d de (X, Y). Nous construisons nos estimateurs avec les observations (X_i, Y_i) pour lesquelles Y_i dépasse un seuil $\mathbf{y}_n$. Ce seuil, $(\mathbf{y}_n)_{n \in \mathbb{N}}$, peut être déterministe, aléatoire ou basé sur les données. Il doit néanmoins dépendre de la taille de l'échantillon $n \geq 1$ de la façon suivante

$$\mathbf{y}_n \xrightarrow{\mathbb{P}} \infty \quad \text{et} \quad N_n \xrightarrow{\mathbb{P}} \infty,$$

avec $N_n := \sum_{i=1}^n \mathbb{I}_{\{Y_i > \mathbf{y}_n\}}$ le nombre d'excès, potentiellement aléatoire. On estime l'indice des valeurs extrêmes $\gamma > 0$ avec un estimateur de type Hill

$$\hat{\gamma}_n = \frac{1}{N_n} \sum_{i=1}^n \log\left(\frac{Y_i}{\mathbf{y}_n} \right) \mathbb{I}_{\{Y_i > \mathbf{y}_n\}}.$$

La fonction C est estimée grâce à une fonction de répartition empirique avec uniquement les observations correspondantes à un excès

$$\widehat{C}_n(x) := \frac{1}{N_n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i>\mathbf{y}_n,\ X_i\leq x\}},\ x\in\mathbb{R}^d.$$

Tout d'abord considérons la normalité asymptotique jointe de $\hat{\gamma}_n$ et $\widehat{C}_n$, c'est-à-dire

$$v_n\begin{pmatrix}\widehat{C}_n(\cdot)-C(\cdot)\\ \hat{\gamma}_n-\gamma\end{pmatrix}\xrightarrow[n\to+\infty]{\mathcal{L}}\mathbb{W}, \tag{5}$$

avec $\mathbb{W}$ une mesure de probabilité Gaussienne sur $L^\infty(\mathbb{R}^d)\times\mathbb{R}$, et $v_n\to+\infty$ une vitesse déterministe.

Théorème 1 *Sous les hypothèses* (3) *et* (4)*, supposons que* $\mathbf{y}_n/y_n\to 1$ *en probabilité pour une suite déterministe* y_n *telle que* $p_n := \bar{F}(y_n)$ *satisfaisant*

$$p_n\to 0,\quad np_n\to+\infty\quad et\quad \sqrt{np_n}^{1+\varepsilon}A\left(\frac{1}{p_n}\right)\to 0\ pour\ un\ \varepsilon>0.$$

Alors, la normalité asymptotique (5) *est vérifiée pour*

$$v_n := \sqrt{np_n}\quad and\quad \mathbb{W}\overset{\mathcal{L}}{=}\begin{pmatrix}B\\ N\end{pmatrix},$$

avec B *un* C*-pont Brownien sur* $\mathbb{R}^d$ *et* N *une variable aléatoire normale centrée de variance* γ^2 *indépendante de* B*.*

Remarque 1 *Il est aussi possible d'établir une version bootstrap de ce théorème. C'est-à-dire d'obtenir la normalité asymptotique jointe de version pondérée des estimateurs* $\hat{\gamma}_n$ *et* $\widehat{C}_n$ *lorsque les données sont fixées. Ce résultat nous permet alors de tester l'adéquation d'un échantillon au modèle des queues proportionnelles en testant si* γ *est constant. Le niveau et la puissance du test d'adéquation ainsi construit ont été étudiés par simulation.*

Nous proposons ensuite un estimateur à noyau pour estimer σ ponctuellement en $x\in\mathbb{R}^d$:

$$\hat{\sigma}_n(x) = n\frac{\sum_{i=1}^n \mathbb{I}_{\{|x-X_i|<h_n\}}\mathbb{I}_{\{Y_i>\mathbf{y}_n\}}}{\sum_{i=1}^n \mathbb{I}_{\{|x-X_i|<h_n\}}\sum_{i=1}^n \mathbb{I}_{\{Y_i>\mathbf{y}_n\}}},$$

avec h_n la largeur de la fenêtre. Dans la suite, nous étudierons les estimateurs uniquement pour des seuils déterministes.

Théorème 2 *Soit* $\mathbf{y}_n\to\infty$ *une suite déterministe et* $p_n := \bar{F}(\mathbf{y}_n)$*. Soit* $h_n\to 0$ *telle que*

$$np_nh_n^d\to+\infty,\quad \sqrt{np_nh_n^d}A\left(\frac{1}{p_n}\right)\to 0.$$

Supposons que f *et* σ *sont toutes deux continues et positives dans un voisinage de* $x\in\mathbb{R}^d$*. Alors, sous l'hypothèse* (3)*, on a*

$$\sqrt{np_nh_n^d}\Big(\hat{\sigma}_n(x)-\sigma(x)\Big)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}\left(0,\frac{\sigma(x)}{f(x)}\right).$$

3 Estimation des quantiles extrêmes

Dans le cadre du seuil déterministe, un estimateur du quantile inspiré par celui de Weissman s'écrirait

$$\hat{q}(\alpha_n) := \mathbf{y}_n \left(\frac{\hat{p}_n}{\alpha_n} \right)^{\hat{\gamma}_n}.$$

avec $\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{Y_i > \mathbf{y}_n\}} = \frac{N_n}{n}$ un estimateur $p_n = \mathbb{P}(Y > \mathbf{y}_n)$ dépendant de la distribution non conditionnelle de Y.
Pour estimer un quantile extrême, développons le lien entre les distributions conditionnelle et non conditionnelle fourni par le modèle des queues proportionnelles. L'hypothèse (1) nous fournit

$$q(\alpha_n \mid x) \sim q\left(\frac{\alpha_n}{\sigma(x)} \right) \quad \text{lorsque } n \to \infty.$$

Ce qui conduit à considérer l'estimateur

$$\hat{q}(\alpha_n|x) := \hat{q}\left(\frac{\alpha_n}{\hat{\sigma}_n(x)} \right) = \mathbf{y}_n \left(\frac{\hat{p}_n \hat{\sigma}_n(x)}{\alpha_n} \right)^{\hat{\gamma}_n}.$$

La normalité asymptotique de cet estimateur peut alors être déduite de celle de $\hat{\gamma}_n$ et de $\hat{\sigma}_n(x)$.

Théorème 3 *Sous les hypothèses des Théorèmes 1 et 2, si* $\sqrt{h_n^d} \log(p_n/\alpha_n) \to +\infty$ *et* $\log(n\alpha_n) = o(\sqrt{np_n})$ *on a alors*

$$\frac{\sqrt{np_n}}{\log(p_n/\alpha_n)} \log\left(\frac{\hat{q}(\alpha_n|x)}{q(\alpha_n|x)} \right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \gamma^2\right).$$

Bibliographie

de Haan, L and Ferreira A. (2006), *Extreme value theory An introduction*, Springer Series in Operations Research and Financial Engineering, Springer, New York.

Einmahl, J.H.J, de Haan, L and Zhou, C. (2016), Statistics of heteroscedastic extremes.*J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78, pp.31-51.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observation. *Journal of American Statistic Association*, 73, pp.812-815.

MODÈLE BAYÉSIEN MULTI-RÉPONSES NON LINÉAIRE À EFFETS MIXTES: APPLICATION À L'ÉVOLUTION DE DEUX BIOMARQUEURS DE L'INFECTION RÉCENTE PAR LE VIH

Charlotte Castel[1,2] & Cécile Sommen[1], & Edouard Chatignoux[1] & Yann Le Strat[1] & Ahmadou Alioum[3,4]

[1] *Santé publique France, Direction Appui, Traitements et Analyses de données (DATA), Saint-Maurice 94415, France*
[2] *Université Paris-Est, Champs-sur-Marne 77420, France*
[3] *Centre Inserm U1219- Bordeaux Population Health, équipe biostatistiques, Bordeaux 33076, France*
[4] *Université de Bordeaux, ISPED, Centre Inserm U1219- Bordeaux Population Health, Bordeaux 33076, France*

charlotte.castel@santepubliquefrance.fr

Résumé. L'épidémie du virus de l'immunodéficience humaine (VIH), découverte il y a 35 ans est toujours active en France. Pour suivre la dynamique de la transmission du VIH et évaluer l'impact des campagnes de prévention, le principal indicateur est l'incidence. Une méthode d'estimation de l'incidence du VIH est basée sur les valeurs des biomarqueurs au moment du diagnostic et de leur dynamique au cours du temps depuis l'infection. L'estimation de l'incidence du VIH à partir des biomarqueurs nécessite au préalable de modéliser leur dynamique depuis l'infection à l'aide de données longitudinales externes. L'objectif des travaux présentés ici est d'estimer la dynamique conjointe de deux biomarqueurs d'infection récente du VIH à partir des données de la cohorte PRIMO. Nous avons modélisé conjointement la dynamique des deux biomarqueurs TM et V3 à l'aide d'un modèle multi-réponses non linéaire à effets mixtes. Les paramètres ont été estimés en utilisant une inférence bayésienne de type Hamiltonien de Monte Carlo. Cette procédure a d'abord été appliquée aux données réelles de la cohorte PRIMO. Dans une étude de simulation, nous avons ensuite évalué les performances de la procédure bayésienne pour estimer les paramètres du modèle multi-réponses non linéaire à effets mixte.

Mots-clés. Modèle non linéaire à effets mixte, Modèle multi-réponses, Inférence de type Hamiltonien, Biomarqueurs du VIH

Abstract. Since the discovery of the human immunodeficiency virus (HIV) 35 years ago, the epidemic is still ongoing in France. To monitor the dynamics of HIV transmission and assess the impact of prevention campaigns, the main indicator is the incidence. One method to estimate the HIV incidence is based on biomarker values at diagnosis and their dynamics over time. Estimating the HIV incidence from biomarkers first requires modeling their dynamics

since infection using external longitudinal data. The objective of the work presented here is to estimate the joint dynamics of two biomarkers from the PRIMO cohort. We thus modeled the dynamics of two biomarkers (TM and V3) using a multi-response nonlinear mixed-effect model. The parameters were estimated using Bayesian Hamiltonian Monte Carlo inference. This procedure was first applied to the real data of the PRIMO cohort. In a simulation study, we then evaluated the performance of the Bayesian procedure for estimating the parameters of multi-response nonlinear mixed-effect models.

Keywords. Nonlinear mixed models, Multi-response model, Hamiltonian Monte Carlo inference, HIV biomarkers

1 Introduction

La transmission du virus de l'immunodéficience humaine (VIH) est toujours une question de santé publique préoccupante dans la plupart des pays. Il est important d'évaluer la dynamique du VIH en estimant son incidence. L'incidence, définie comme le nombre de nouvelles infections au cours d'une période donnée, est un indicateur épidémiologique majeur pour surveiller la dynamique d'une maladie et évaluer l'impact des campagnes de prévention. Depuis plusieurs années, plusieurs méthodes statistiques ont été développées pour estimer l'incidence du VIH selon les différents types de données collectées: cohortes, enquêtes transversales ou systèmes de notification. En France, le système de notification obligatoire des diagnostics de séropositivité VIH a été mis en place en mars 2003 par Santé publique France. Associée aux données de surveillance des nouveaux diagnostics VIH, une surveillance biologique a également été mise en place pour recueillir deux biomarqueurs appelés TM et V3, qui permettent la distinction entre les infections récentes (moins de 6 mois en moyenne) et celles plus anciennes (plus de 6 mois en moyenne).
Sur la base de ces données de notification du VIH, des méthodes d'estimation d'incidence ont été proposées, basées sur les valeurs des biomarqueurs au moment du diagnostic et sur la connaissance de leur dynamique depuis l'infection . [Sommen et al., 2011, Le Vu et al., 2010]. Au préalable, la dynamique des deux biomarqueurs doit d'abord être estimée sur la base de données longitudinales externes, en utilisant par exemple des modèles non linéaire à effets mixtes. Ces types de modèles sont très largement utilisés pour modéliser la dynamique de biomarqueurs [Tuerlinckx et al., 2006, Gad and El Kholy, 2012].
Dans ce travail, nous modélisons conjointement la dynamique des biomarqueurs TM et V3 en utilisant un modèle non linéaire à effets mixtes. Nous avons considéré un effet aléatoire pour chaque biomarqueur et une corrélation des effets aléatoires pour prendre en compte la corrélation des deux biomarqueurs. Les paramètres ont été estimés en utilisant une inférence de type Hamiltonien de Monte Carlo (HMC).
À notre connaissance, ce travail est la première tentative d'étudier conjointement la dynamique

de deux biomarqueurs dans un modèle bayésien non linéaire à effets mixtes. Pour modéliser la dynamique de ces marqueurs biologiques, nous avons utilisé les données de la cohorte ouverte PRIMO-ANRS C06 [Desquilbet et al., 2002]. Cette cohorte comprend 298 volontaires infectés avec le VIH au stade d'infection primaire ayant une date d'infection connue ou estimée. Un suivi longitudinal des patients a eu lieu pendant toute la période de la cohorte avec la collecte des valeurs des biomarqueurs TM et V3.
Nous avons d'abord estimé les paramètres du modèle à partir des données de la cohorte PRIMO-ANRS C06, puis simulé des données les plus proches possibles des données réelles afin de valider la procédure d'estimation et le choix des lois *a priori*. La section 2 décrit le modèle bayésien. Les données et les résultats de la cohorte PRIMO ANRS C06 sont présentés dans les sections 3 et 4, respectivement. La section 5 présente la validation de la procédure avec l'étude de simulation. Enfin, nous discutons des avantages de ce modèle et de son utilité pour estimer l'incidence.

2 Modèle

2.1 Spécification du modèle

Nous proposons une version simplifiée du modèle développé par Sommen et al. [Sommen et al., 2011]. Soient $\mathbf{Y_i^1} = \left(Y_{i1}^1, ..., Y_{i,j}^1, ..., Y_{in_i}^1\right)$ et $\mathbf{Y_i^2} = \left(Y_{i1}^2, ..., Y_{ij}^2, ..., Y_{in_i}^2\right)$ les valeurs des biomarqueurs TM et V3 respectivement, où j représente la j^{eme} $(j = 1, ..., n_i)$ mesure pour un individu i au temps calendaire t_{ij}, et $\mathbf{Y_i} = (\mathbf{Y_i^1}, \mathbf{Y_i^2})^T$. On suppose que les deux biomarqueurs TM et V3 sont mesurés au même temps calendaire t_{ij}. Soit u_i le temps d'infection connu (ou estimé par le clinicien) pour l'individu i. Nous considérons le modèle multi-réponses non linéaire à effets mixtes suivant pour les observations Y_{ij}^1 et Y_{ij}^2:

$$Y_{ij}^1 = g_{\mathbf{b^1}}(t_{ij} - u_i, a_i^1) + \varepsilon_{ij}^1$$
$$Y_{ij}^2 = g_{\mathbf{b^2}}(t_{ij} - u_i, a_i^2) + \varepsilon_{ij}^2$$

Les fonctions $g_{\mathbf{b^1}}(.)$ et $g_{\mathbf{b^2}}(.)$ représentent la dynamique moyenne de la concentration des biomarqueurs TM et V3 depuis le temps d'infection u_i. Ces fonctions dépendent de deux vecteurs de paramètres: $\mathbf{b^1}$ pour le biomarqueur TM et $\mathbf{b^2}$ pour le biomarqueur V3, et d'effets aléatoires $\mathbf{a_i} = (\mathbf{a_i^1}, \mathbf{a_i^2})$ supposés suivre une distribution gaussienne avec une moyenne de zéro et une matrice de variance-covariance $\Sigma_a = \begin{pmatrix} \sigma_{a^1}^2 & \sigma_{a^1a^2} \\ \sigma_{a^1a^2} & \sigma_{a^2}^2 \end{pmatrix}$. Nous considérons un unique effet aléatoire par biomarqueur. L'effet aléatoire permet de prendre en compte la variabilité individuelle de l'évolution de la concentration des biomarqueurs TM et V3. La corrélation entre les deux biomarqueurs est prise en compte à travers la corrélation de leurs effets aléatoires. Les erreurs de mesure $\varepsilon_{ij} = (\varepsilon_{ij}^1, \varepsilon_{ij}^2)$ sont indépendantes et supposées suivre une distribution gaussienne de moyenne zéro avec des variances respectives $\sigma_{\varepsilon^1}^2$ et $\sigma_{\varepsilon^2}^2$. De plus, nous supposons que $\mathbf{a_i}$ et ε_{ij}

sont indépendants. Une approche bayésienne a été proposée pour l'estimation des paramètres et la validation du modèle. Pour la fonction $g_{\mathbf{b^k}}(.)$, nous avons considéré trois familles de fonctions sigmoïdes. Ces familles de fonctions dépendent de paramètres fixes $\mathbf{b^k} = (b_1^k, b_2^k, b_3^k)^T$, $k = 1$ pour le biomarqueur TM et $k = 2$ pour le biomarqueur V3. Elles dépendent aussi d'effets aléatoires $\mathbf{a} = (a_1, a_2, a_3)^T$, mais comme indiqué ci-dessus, nous considérons un unique effet aléatoire par fonction. Cela se traduit par le fait qu'un seul effet aléatoire a_1, a_2 ou a_3 sera placé sur l'un des paramètre fixe b_1, b_2 ou b_3.

Dans ce qui suit, nous avons considéré la même famille de fonctions $g_{\mathbf{b}}^k$ pour les deux biomarqueurs afin d'interpréter les résultats de la même manière. Le choix de la place de l'effet aléatoire sur les paramètres fixes b_1, b_2 ou b_3 (a_1, a_2, a_3) se fait en examinant toutes les possibilités pour chacune des 3 fonctions sigmoïdes. Nous avons testé les 3 familles de fonctions et pour chaque famille de fonctions nous avons testé les 3 places possibles de l'effet aléatoire, ceci représentant 9 modèles différents.

2.2 Spécifications des lois *a priori*

Dans le modèle décrit dans la section précédente, les paramètres à estimer sont les vecteurs $\mathbf{b^1}$, $\mathbf{b^2}$, les composantes de la matrice de variance-covariance Σ_a des effets aléatoires, et les variances $\sigma^2_{\varepsilon^1}$ et $\sigma^2_{\varepsilon^2}$ des erreurs de mesure. Pour tous ces paramètres nous devons spécifier la distribution des lois *a priori*. Comme nous n'avions pas d'information sur ces lois *a priori* issue de précédentes études, nous avons utilisé des lois *a priori* faiblement informatives. Nous avons suivi les recommandations de Gelman et al. [Gelman et al., 2013a] pour choisir une distribution normale $\mathscr{N}(0;10000)$ comme *a priori* faiblement informatif pour les paramètres fixes et une distribution uniforme avec une grande borne supérieure $\mathscr{U}(0;10000)$ pour les paramètres de variances et de covariances.

2.3 Estimation et sélection du modèle

Les paramètres du modèle se trouvent dans un espace multidimensionnel, ce qui a orienté notre choix vers une méthode de type Hamiltonien de Monte-Carlo (HMC) plutôt que vers une méthode MCMC plus classique. A la différence du MCMC classique qui propose une nouvelle position des paramètres toujours centrée sur la position actuelle, le HMC propose une distribution des paramètres qui change en fonction de la position actuelle. Le HMC détermine la direction dans laquelle la distribution *a posteriori* augmente, appelée son gradient, et déforme la distribution des paramètres vers le gradient.[Kruschke, 2015]. Nous avons utilisé le critère WAIC [Watanabe, 2009, 2010] pour comparer les différents modèles que nous avons testés. Comme indiqué dans la section ci-dessus, l'objectif était de choisir un modèle qui optimisait le mieux la place des effets aléatoires. Nous avons utilisé ce critère pour sélectionner le modèle optimal. Ce critère est interprété de la même manière que le critère AIC. Le modèle final est celui correspondant à la valeur de WAIC la plus faible.

3 Données de la cohorte PRIMO ANRS C06

La cohorte ouverte PRIMO-ANRS C06 a recruté 298 volontaires infectés par le VIH entre novembre 1996 et septembre 2007 dans 66 hôpitaux français [Desquilbet et al., 2002]. Après avoir donné leur consentement écrit, les patients ont été inclus dans la cohorte s'ils étaient au stade d'infection primaire du VIH avec ou sans symptômes. De plus, les patients devaient être naïfs d'antirétroviraux pour être inclus dans la cohorte.
Après leur inclusion, les patients ont été suivis cliniquement et biologiquement au premier mois, au troisième mois, au sixième mois, puis tous les six mois. À chaque visite, la concentration des anticorps TM et V3 a été mesurée.
Une première étude pour estimer la dynamique des biomarqueurs à partir de cette cohorte dans un cadre fréquentiste a déjà été réalisée par Sommen et al. [Sommen et al., 2011] et a impliqué 248 patients et 585 observations avec un suivi maximum de 550 jours depuis l'infection. Pour l'analyse actuelle, 272 patients et 812 mesures ont été inclus, avec un suivi maximal de 975 jours depuis l'infection après avoir exclu les patients sans aucune mesure de marqueur et ceux avec une concentration d'anticorps plate au fil du temps (c'est-à-dire non progresseurs). Cette nouvelle analyse a donc été réalisée avec plus de patients, plus d'observations et un temps de suivi plus long. Parmi ces 272 patients, 134 avaient une unique mesure de marqueur, 8 en avaient deux, 12 en avaient trois, 47 en avaient quatre, 13 en avaient cinq, 33 en avaient six et 25 avaient sept mesures de marqueurs avec un suivi maximal de 975 jours depuis l'infection. Les biomarqueurs TM et V3 sont censurés à la valeur 70 en raison de la saturation du signal lorsqu'il y a une très forte réaction positive des biomarqueurs.

4 Résultats

4.1 Hypothèses

Les estimations du modèle sur données réelles impliquaient quatre chaînes de 20000 itérations, après exclusion d'un échauffement de la même taille. Pour vérifier que les quatre chaînes convergeaient vers les mêmes valeurs de paramètres à partir de valeurs initiales très différentes, nous avons utilisé le critère de Gelman-Rubin [Gelman et al., 2013b]. Toutes les estimations ont été réalisées à l'aide du logiciel R version 3.5.1 et notamment avec le package brms créé par Burkner [Burkner, 2017].

4.2 Résultats sur les données PRIMO

L'estimation a été réalisée à partir de mesures répétées des biomarqueurs TM et V3 de la cohorte PRIMO-ANRS C06. Nous avons testé 3 familles de fonctions sigmoïdes en faisant varier la place de l'effet aléatoire sur le paramètre b_1, b_2 ou b_3 (a_1, a_2, a_3). Nous avons sélectionné le modèle final parmi les 9 modèles testés en utilisant le critère WAIC.

4.3 Validation du modèle sur données PRIMO

Pour évaluer la qualité de l'ajustement du modèle final que nous avons sélectionné, nous avons comparé les trajectoires individuelles observées dans la cohorte PRIMO avec celles prédites par le modèle pour tous les individus. Nous avons estimé les effets aléatoires de chaque individu par le mode *a posteriori*, c'est à dire en connaissant les données.
La Figure 1 présente les trajectoires individuelles observées dans la cohorte PRIMO avec celles prédites par le modèle pour neuf individus choisis avec sept observations. Les neuf individus ont été choisis aléatoirement. Nous avons pris les neuf premiers individus de la cohorte qui avaient sept mesures de marqueurs. Les trajectoires prédites sont proches des trajectoires observées pour les neuf individus. Les trajectoires prédites sont proches des trajectoires observées pour tous les autres individus sauf pour quatre individus. Ces quatre individus n'ont pas convergé car leurs valeurs de biomarqueurs n'avaient pas une dynamique classique de sigmoïde.

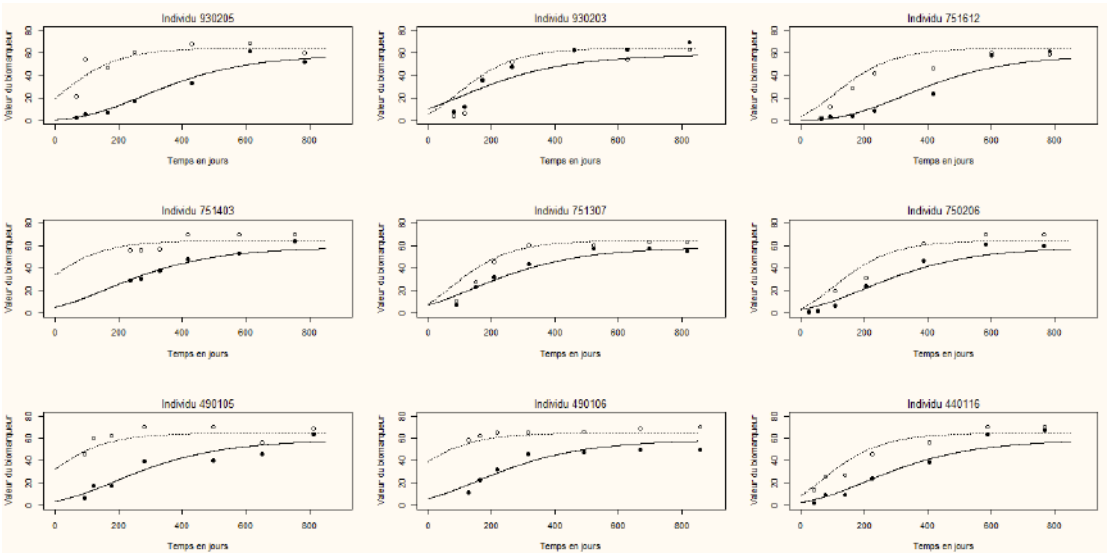


Figure 1: Valeurs des biomarqueurs TM et V3 observés (cercle plein pour TM et cercle vide pour V3) et trajectoires prédites (courbe pleine pour TM et pointillé pour V3) pour 9 sujets de la cohorte PRIMO-ANRS C06

5 Etude de simulation

5.1 Données simulées

L'étude de simulation visait à créer un ensemble de données réalistes qui était aussi proche que possible de la dynamique des biomarqueurs TM et V3 observée dans la cohorte PRIMO

ANRS Cohorte C06 afin de vérifier que le modèle ne présentait aucun problème de calcul et qu'il reflétait correctement le modèle mathématique choisi. Le but de la simulation consistait également à vérifier la correcte spécification des lois *a priori* et valider la procédure d'estimation bayésienne. L'étude de simulation était basée sur 200 bases de données simulées avec chacune 272 individus comme dans la cohorte PRIMO.

5.2 Résultats des simulations

Pour évaluer le comportement des estimations bayésiennes, nous avons effectué 200 simulations et calculé les indicateurs classiques tels que l'erreur quadratique moyenne empirique (RMSE), la moyenne empirique, le biais relatif absolu (ARB) et le taux de couverture (CR) de l'intervalle de crédibilité à 95% pour chacun des paramètres estimés. Nous avons utilisé quatre chaînes pour chacune des 200 simulations, et à chaque fois, les chaînes convergeaient vers les mêmes valeurs de paramètres.
Pour l'étude de simulation, le biais relatif absolu est compris entre 0% et 8% , les RMSE se situent entre 0,11 et 2,1 et le taux de couverture est compris entre 90% et 98% . De même que pour les données de la cohorte PRIMO, nous avons évalué la qualité de l'ajustement de notre modèle sur les données simulées. Nous avons comparé les trajectoires individuelles observées dans la simulation et celles prédites par le modèle pour tous les individus, en estimant les effets aléatoires individuels par le mode *a posteriori*. Les trajectoires prédites sont proches des trajectoires observées pour tous les individus des 200 bases de données simulées.

6 Discussion

Le modèle bayésien multi-réponses non linéaire à effets mixtes présenté dans ce travail est une alternative à l'approche fréquentiste pour estimer la dynamique conjointe de deux biomarqueurs. Les avantages de cette méthode sont qu'elle estime précisément les paramètres de modèles multi-réponses non linéaires à effets mixtes, en tenant compte de la corrélation entre les deux biomarqueurs, ce qui n'est pas le cas pour les autres méthodes actuellement proposées pour ce type de modèle [Lachos et al., 2013, Bazzoli et al., 2010, Lavielle and Mentré, 2007]. De plus, ce type de modèle ne repose pas sur des hypothèses fortes. Un second avantage de cette méthode est le temps de calcul raisonnable (2 heures). Une limite de ce modèle est la difficulté de choisir les lois *a priori*, bien que cela soit le cas pour toute estimation bayésienne. Cependant, il est important d'observer que le modèle converge toujours malgré la spécification des lois *a priori* non informative.
L'étude de simulation a donné de bons résultats, puisque nous avons obtenu des valeurs faibles de biais absolu relatif, d'erreur quadratique moyenne et à l'inverse nous avons obtenu des taux de couverture corrects, ceci pour tous les paramètres. Si nous supposons que les données réelles sont proches des données simulées, on s'attend alors à obtenir des estimations "aussi fiables" avec le modèle et les spécifications des lois *a priori*. Les résultats obtenus sur les données

simulées confirment que le modèle bayésien n'a pas de problème de calcul et que les lois *a priori* choisies sont pertinentes. On peut donc dire que la procédure d'estimation bayésienne de type Hamiltonien de Monte Carlo fonctionne correctement pour estimer ce type de modèle. Cette méthode nous permet ainsi d'estimer précisément les paramètres d'un modèle multi-réponses non linéaire à effets mixtes. Ce modèle permet de modéliser conjointement la dynamique des deux biomarqueurs d'une manière précise et rapide. L'approche n'est pas basée sur des hypothèses fortes spécifiques aux valeurs des biomarqueurs TM et V3. En conséquence, ce travail pourrait potentiellement fournir un cadre pour appliquer ce modèle à d'autres biomarqueurs en modifiant la fonction g_b, qui modélise l'évolution des anticorps.
Dans la notification obligatoire du VIH, les biomarqueurs TM et V3 sont collectés pour chaque individu au moment du diagnostic. En termes de perspectives, cette modélisation sera utilisée pour estimer l'incidence du VIH sur la base des valeurs des biomarqueurs TM et V3 fournies par la notification obligatoire du VIH au moment du diagnostic.

References

C. Bazzoli, S. Retout, and F. Mentré. Design evaluation and optimisation in multiple response nonlinear mixed effect models: Pfim 3.0. *Computer methods and programs in biomedicine*, 98, 2010.

P.C. Burkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 2017.

L. Desquilbet, C. Deveau, C. Goujard, J.B. Hubert, J. Derouineau, L. Meyer, and the PRIMO Cohort Study Group. Increase in at-risk sexual behaviour among HIV-1- infected patients followed in the french primo cohort. *AIDS*, 16:2329–2333, 2002.

A.M. Gad and R.B. El Kholy. Generalized linear mixed models for longitudinal data. *International Journal of Probability and Statistics*, 1, 2012.

A. Gelman, J. Carlinb, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, third edition*. CRC Press, London, 2013a.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2013b.

J.K. Kruschke. *Doing bayesian data analysis*. Elsevier, London, 2015.

V.H. Lachos, L. M. Castro, and D.D Dey. Bayesian inference in nonlinear mixed-effects models using normal independent distributions. *Computational Statistics and Data Analysis*, 64, 2013.

M. Lavielle and F. Mentré. Estimation of population pharmacokinetic parameters of saquinavir in HIV patients with the monolix software. *Journal of Pharmacokinetics and Pharmacodynamics*, 34, 2007.

S. Le Vu, Y. Le Strat, F. Barin, J. Pillonel, F. Cazein, V. Bousquet, S. Brunet, D. Thierry, C. Semaille, L. Meyer, and J.C. Desenclos. Population-based HIV-1 incidence in france, 2003–08:a modelling analysis. *Lancet infectious diseases*, 10, 2010.

C. Sommen, D. Commenges, S. Le Vu, L. Meyer, and A. Alioum. Estimation of the distribution of infection times using longitudinal serological markers of HIV: implications for the estimation of HIV incidence. *Biometrics*, 67(11), 2011.

F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. De Boeck. Statistical inference in generalized linear mixed models: A review. *The British Psychological Society*, 59:225–255, 2006.

S. Watanabe. Algebraic geometry and statistical learning theory. *Cambridge University Press*, 2009.

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11: 3571–3594, 2010.

Functional Peaks-over-threshold Analysis and its Applications in Environment

Raphaël de Fondeville [1] & Anthony C. Davison[2]

[1] *raphael.de-fondeville@epfl.ch - Swiss Data Science Center, INN 218, Station 14, 1015 Lausanne, Switzerland*

[2] *anthony.davison@epfl.ch - Chair of Statistics, École polytechnique fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland*

Résumé. Les techniques de quantification du risque naturel se sont largement démocratisées ces dernières années. Elles sont encore toutefois majoritairement limitées à la simple utilisation de catalogues d'évènements historiques, dont la taille excède rarement 40 à 50 ans, ainsi qu'à l'exploitation de modèles numériques, impliquant de lourds calculs tout en n'étant pas fiable pour l'extrapolation. La théorie des valeurs extrêmes définit les principes d'analyse statistiques nécessaires à l'estimation de la fréquence d'évènements rares tout en donnant un cadre formel pour extrapoler au-delà des niveaux historiques d'intensité. Toutefois son application s'est jusque-là principalement limitée au cadre univarié. Ainsi, une majorité des études traitant du risque naturel ont négligé sa nature spatio-temporelle.

Dans cette présentation, nous introduisons une extension de l'analyse de dépassements de seuil au cadre fonctionnel, dans lequel il est possible de caractériser un évènement extrême complexe à travers une notion généralisée d'excès, et décrivons ensuite la limite de leur queue de distribution, appelée processus de r-Pareto généralisé. Nous présentons un modèle dérivé de fonctions aléatoires log-Gaussiennes qui utilise les structures classiques de covariance pour caractériser la dépendance extrémale. Ensuite, nous décrivons un générateur stochastique d'évènements extrêmes, capable de quantifier la récurrence d'évènements passés ainsi que d'en générer des nouveaux dont l'intensité va au-delà des niveaux historiques. La méthodologie est ensuite appliquée à plusieurs risques naturels tels que les tempêtes et la pluie.

Mots-clés. Analyse de dépassements de seuil, Extrêmes spatio-temporels, Processus de r-Pareto généralisé, Risque naturel.

Abstract. Estimating the risk of single occurrences of natural hazards has become important in recent decades, but up until now it has been largely limited to re-using catalogues of historical events, which usually do not exceed 40 to 50 years in length, and to numerical models, which require heavy computation and are often unreliable for extrapolation. Extreme value theory provides statistical methods for estimating the frequency of past extreme events as well as for extrapolating beyond observed severities, but it has mostly been focused on studying univariate quantities. Consequently the majority of its applications to natural hazards have neglected their spatio-temporal characteristics.

We present an extension of peaks-over-threshold analysis to functions which allows one to define complex extreme events as special types of exceedances, and then obtain their limit tail distribution, namely the generalized r-Pareto process. We focus on a specific model based on log-Gaussian random functions using classical covariance structures to characterize extremal dependence. Then, we describe a stochastic weather generator for extreme events, capable of quantifying the recurrence of past events as well as generating completely new ones. The methodology is applied to several natural hazards such as windstorms and rainfall.

Keywords. Generalized r-Pareto process, natural hazards, peaks-over-threshold analysis, spatio-temporal extremes.

1 Extended summary

Extreme Value Theory (EVT) provides a theoretical framework to describe and model tails of statistical distributions within which estimating the frequency of past extreme events as well as to extrapolating beyond observed severities is possible. These have been extensively studied in a univariate framework (Fisher and Tippett, 1928; Gnedenko, 1943; Davison and Smith, 1990) especially for independent identically distributed replicates, and applications have been developed in fields such as finance, insurance, hydrology and telecommunications (Hosking and Wallis, 1987; Katz et al., 2002; Embrechts et al., 1997). Due to recent extreme events, there has been a surge of interest in environmental applications, motivated by the necessity to better understand the impact of global warming. Floods, windstorms, heatwaves have a complex spatio-temporal structure that cannot be modelled using univariate extreme value theory.

Max-stable processes (de Haan and Ferreira, 2006, Section 9.2), which provide a functional extension of the generalized extreme value distribution (Coles, 2001, p.47-48), have successfully been used to study the extremal behaviour of monthly and annual maxima, but applications have been limited due to the mathematical and computational complexity of such models (Huser and Davison, 2013). Also, the study of maxima discards a fair amount of information, making detection of mixtures in tail behaviour very difficult. For example, in some regions, rainfall can be divided into two classes: convective rain, which is local and marginally very intense, and cyclonic spells generating larger spatial accumulations of water but with lower local intensities. These phenomena are driven by different independent weather conditions that may both cause severe floods and their tail marginal distribution and spatio-temporal structure are likely to differ. With block maxima, marginally intense events naturally dominate and thus impose a focus on convective rainfall, while disregarding potential extreme cyclonic events. For risk mitigation, studying extremes of different natures is crucial, and max-stable processes are inappropriate

for modelling such complex phenomena, since taking maxima largely eliminates certain types of events.

Univariate peaks-over-thresholds analysis, associated to the generalized Pareto distribution, define extreme events as exceedances over a threshold. In this context, reduction of multivariate datasets to univariate structural variables, such as $\max(X_1, X_2)$ or $X_1^2 + X_2^2$, on which generalized Pareto distributions are fitted (Coles and Tawn, 1994), is common to study complex multivariate extreme events. However, this approach does not give insight on the combination of events yielding an exceedance and is hindered by the fact that different univariate summaries may lead to different tail behaviour. One way to understand these differences is to suppose that the observations are generated by an underlying mixture of generative processes, which are disentangled by computing these univariate summaries. Thus if the summary captures only one of these processes, for instance only cyclonic rain, it is not surprising that we obtain different tail behaviours. Functional peaks-over-threshold analysis generalizes this methodology for a better understanding of the underlying dependence structure and gives a theoretical foundation to detect mixtures of tail behaviour through different definitions of exceedances tailored to the type of extreme events of interest.

In univariate extreme value theory, the generalized Pareto distribution gives a unified framework to describe directly the tail decay of the original data, and encompasses the Weibull, Gumbel and Fréchet tail decay regimes. This work provides a similar unified formulation for functional peaks-over-threshold analysis under the assumption that the process has the same tail decay over its domain. In this context, we extend Dombry and Ribatet (2015) by introducing the generalized r-Pareto process, allowing more flexible excess definitions and generalized Pareto tail margins. The generalized r-Pareto process is the only limit of exceedances of a properly rescaled regularly varying process and for some specific definitions of exceedance, it can be factorized to enable simulation of events with a fixed intensity, i.e., events for which the risk measure equals a pre-determined return level.

We first review classical results for univariate extremes and introduces functional peaks-over-threshold analysis. We present convergence results for the three possible regimes of tail decay, under a generalized regular variation hypothesis, i.e., for a stochastic process X, we assume that there exist a tail index $\xi \in \mathbb{R}$ and sequences of functions $a_n > 0$ and b_n such that

$$n\Pr\left\{\left(1 + \xi\frac{X - b_n}{a_n}\right)^{1/\xi} \in \cdot\right\} \to \Lambda(\cdot), \quad n \to \infty,$$

where Λ is a non-zero measure on the space of non-negative continuous functions. We then introduce the class of generalized r-Pareto process, characerized by

$$P = r(a)\xi^{-1}R^{\xi}\frac{W}{\mathrm{r}(W)} + b - \xi 1a,$$

where R is a unit Pareto random variable, W is a stochastic process on the space of continuous functions with unit norm and $a > 0$ and b are continuous functions. For linear risk functionals, we prove that generalized r-Pareto processes are the only limit of increasingly large r-exceedances, i.e., events $\{r(X) > u\}$ with increasing threshold $u \in \mathbb{R}$. The previous result is then applied to develop of stochastic weather generator of windstorms over Europe. Finally we illustrate the importance of risk definition when studying the risk of flooding in the city of Zurich.

References

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Coles, S. G. and Tawn, J. A. (1994). Statistical Methods for Multivariate to Structural Design Extremes: an Application to Structural Design. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):1–48.

Davison, A. C. and Smith, R. L. (1990). Models for Exceedances over High Thresholds (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York, USA.

Dombry, C. and Ribatet, M. (2015). Functional Regular Variations, Pareto Processes and Peaks Over Thresholds. *Statistics and Its Interface*, 8(1):9–17.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

Ferreira, A. and de Haan, L. (2014). The Generalized Pareto Process; with a View Towards Application and Simulation. *Bernoulli*, 20(4):1717–1737.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting Forms of the Frequency Fistribution of the Largest or Smallest Member of a Sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Gnedenko, B. (1943). Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Annals of Mathematics*, 44(3):423–453.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and Quantile Estimation for Generalized Pareto Distribution. *Technometrics*, 29(3):339–349.

Huser, R. and Davison, A. C. (2013). Composite Likelihood Estimation for the Brown–Resnick Process. *Biometrika*, 100(2):511–518.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of Extremes in Climatology. *Advances in Water Resources*, 25(8-12):1287–1304.

Klüppelberg, C. and Resnick, S. I. (2008). The Pareto Copula, Aggregation of Risks, and the Emperor's Socks. *Journal of Applied Probability*, 45(1):67–84.

Filtre de Kalman, application à la prévision en ligne de consommation d'électricité

Joseph de Vilmarest[1], Yannig Goude[2], Thi Thu Huong Hoang[3] & Olivier Wintenberger[4]

[1] *LPSM, Sorbonne Université and EDF R&D,* joseph.de_vilmarest@upmc.fr
[2] *EDF R&D,* yannig.goude@edf.fr
[3] *EDF R&D,* thi-thu-huong.hoang@edf.fr
[4] *LPSM, Sorbonne Université,* olivier.wintenberger@upmc.fr

Résumé. L'électricité étant une énergie qui se stocke difficilement, un enjeu crucial chez EDF est la prévision de la consommation électrique pour produire au plus proche de ce qui est consommé. L'évolution du comportement des consommateurs ainsi que l'ouverture à la concurrence du marché électrique rend les séries temporelles étudiées non stationnaires, motivant la mise au point de modèles qui évoluent au cours du temps. Nous nous intéressons à un modèle espace-état linéaire gaussien, résolu par les célèbres formules récursives de Kalman. Nous étudions l'algorithme dans le cas statique (modèle linéaire gaussien), pour lequel on présente une borne sur le risque cumulé, ouvrant la voie à une analyse du cas dynamique. On applique le filtre de Kalman au problème de la prévision de consommation pendant la rupture du confinement de 2020. Dans le cadre de notre application, de nombreux modèles utilisés en opérationnel reposent sur les modèles additifs généralisés (GAM). En figeant les effets non linéaires du modèle GAM, nous mettons en évidence un gain par filtrage de Kalman sur le modèle linéaire obtenu.

Mots-clés. Filtre de Kalman, consommation électrique

Abstract. As electricity is difficult to store, a crucial issue at EDF is to predict the electricity load in order to produce as close as possible to what is consumed. Changes in the consumer's behavior along with the introduction of competition in the electricity market make the demand non stationary. Thus there is a need for adaptive models of prediction. We consider a linear gaussian state-space model yielding the well-known Kalman recursive formulae. We analyse the Kalman Filter in the static case (linear gaussian model), where we derive a bound on the cumulative risk holding with high probability, paving the way to the analysis of dynamic settings. We experiment the Kalman Filter in the context of electricity load forecasting during the lockdown of Spring 2020. In this application, operational models rely on generalized additive models (GAM). We freeze non-linear transforms and we show a predictive gain using Kalman filtering on the linear model obtained.

Keywords. Kalman Filter, electricity consumption

Algorithm 1: Filtre de Kalman

1. *Initialisation*: $\hat{\theta}_1 \in \mathbb{R}^d$, P_1 positive semi-définie.
2. *Itération*: pour tout t,
   (a) Prévision: $\hat{y}_t = \hat{\theta}_t^\top x_t$.
   (b) Mise à jour de $\hat{\theta}$: $\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{P_t x_t}{x_t^\top P_t x_t + \sigma^2}(y_t - \hat{\theta}_t^\top x_t)$.
   (c) Mise à jour de P: $P_{t+1} = P_t - \frac{P_t x_t x_t^\top P_t}{x_t^\top P_t x_t + \sigma^2} + Q$.

1 Le modèle espace-état

On se place dans un cadre en ligne, où l'on prévoit de façon itérative $y_t \in \mathbb{R}$ à l'aide de variables explicatives $x_t \in \mathbb{R}^d$ ainsi que des valeurs passées $(x_s, y_s)_{s<t}$. L'objectif est de produire une prévision $\hat{y}_t$ à chaque étape en vue de minimiser l'erreur quadratique $\sum_t (y_t - \hat{y}_t)^2$.

La régression linéaire est un modèle statique qui permet de prévoir y_t sachant x_t. Cela consiste à estimer y_t par $\theta^\top x_t$ pour un certain $\theta \in \mathbb{R}^d$. Le paramètre θ peut être obtenu, entre autres, par moindres carrés ordinaires qui donnent l'optimum de la perte empirique, ou bien par régression LASSO ou Ridge qui régularisent les moindres carrés ordinaires.

Dans un cadre dynamique, on considère le modèle espace-état suivant:

$$\begin{aligned} \text{Equation d'espace:} \quad & y_t = \theta_t^\top x_t + \varepsilon_t, \\ \text{Equation d'état:} \quad & \theta_{t+1} = \theta_t + \eta_t, \end{aligned}$$

où $(\varepsilon_t)_t$ et $(\eta_t)_t$ sont des bruits blancs gaussiens centrés de variance/covariance respectives σ^2 et Q. Lorsque le modèle est bien spécifié, et si les paramètres σ^2, Q ainsi qu'un prior $\theta_1 \sim \mathcal{N}(\hat{\theta}_1, P_1)$ sont connus, le filtre de Kalman (Algorithme 1) permet d'obtenir de façon récursive les valeurs de

$$\hat{\theta}_t = \mathbb{E}[\theta_t \mid (x_s, y_s)_{s<t}], \qquad P_t = \mathbb{E}[(\theta_t - \hat{\theta}_t)(\theta_t - \hat{\theta}_t)^T \mid (x_s, y_s)_{s<t}].$$

On en déduit alors $y_t \sim \mathcal{N}(\hat{\theta}_t^\top x_t, \sigma^2 + x_t^\top P_t x_t)$. Voir par exemple Durbin et Koopman (2012) pour la preuve des formules de Kalman.

2 Cas statique

On étudie le cas dégénéré où $Q = 0$ (de Vilmarest et Wintenberger, 2020). Il est à noter que l'estimateur $\hat{\theta}_t$ renvoyé par l'Algorithme 1 ne dépend alors que de $\hat{\theta}_1, P_1/\sigma^2$ et l'on

utilise donc $\sigma^2 = 1$. Le filtre de Kalman est équivalent à une régression ridge pour un paramètre décroissant de régularisation:

Proposition 1. *Pour toute séquence (x_t, y_t), si $\hat{\theta}_1 \in \mathbb{R}^d, P_1 \succ 0$, l'algorithme 1 donne*

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \left(\frac{1}{2} \sum_{s=1}^{t-1} (y_s - \theta^\top x_s)^2 + \frac{1}{2} (\theta - \hat{\theta}_1)^\top P_1^{-1} (\theta - \hat{\theta}_1) \right), \qquad t \geq 1 \,.$$

Ainsi, le poids de la régularisation L^2 décroît en $1/t$. Dans le cadre adversarial, une borne de regret a été obtenue par Cesa-Bianchi et Lugosi (2006), mais elle nécessite de borner y_t. On se place dans un cadre stochastique qui consiste à minimiser le risque plutôt que la perte, et l'on suppose les données indépendantes et de même loi:

Hypothèse 1. *Les observations $(x_t, y_t)_t$ sont des copies indépendantes de (x, y), $\mathbb{E}[xx^\top]$ est positive définie et x est borné presque sûrement par D_x.*

Sous l'hypothèse 1, on définit le risque $L(\theta) = \mathbb{E}\left[(y - \theta^\top x)^2\right]$. Pour obtenir un problème bien défini, on suppose:

Hypothèse 2. *Il existe $\theta^\star \in \mathbb{R}^d$ tel que $L(\theta^\star) = \inf_{\theta \in \mathbb{R}^d} L(\theta)$.*

Enfin, on n'a pas besoin de supposer que $\mathcal{L}(y \mid x) = \mathcal{N}(\theta^{\star\top} x, \sigma^2)$ (modèle bien spécifié), mais on suppose qu'on n'en est pas trop loin: d'une part y est sous-gaussien conditionnellement à x, et d'autre part l'erreur d'approximation est bornée:

Hypothèse 3. *La distribution de (x, y) vérifie l'existence de*

- *$\sigma_{\text{sg}}^2 > 0$ tel que pour tout $s \in \mathbb{R}$, $\mathbb{E}\left[e^{s(y - \mathbb{E}[y|x])} \mid x\right] \leq e^{\frac{\sigma_{\text{sg}}^2 s^2}{2}}$ p.s.*
- *$D_{\text{app}} \geq 0$ tel que $|\mathbb{E}[y \mid x] - \theta^{\star\top} x| \leq D_{\text{app}}$ p.s.*

On obtient alors une forte propriété de convergence de l'estimateur $\hat{\theta}_t$ vers $\theta^\star$. Avec grande probabilité, l'algorithme est piégé dans une région locale autour de l'optimum en un temps pour lequel on peut obtenir une borne non-asymptotique.

Proposition 2. *Si les hypothèses 1, 2, 3 sont vérifiées, alors pour tout $\varepsilon, \delta > 0$, on a $\tau(\varepsilon, \delta) = O\left(\varepsilon^{-1} \ln \delta^{-1} \ln(\varepsilon^{-1} \ln \delta^{-1})\right)$ tel qu'on a simultanément*

$$\forall t > \tau(\varepsilon, \delta), \qquad \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon \,,$$

avec probabilité supérieure à $1 - \delta$.

Grâce à cette convergence de l'algorithme, on décompose en deux le risque cumulé. Jusqu'à $\tau(\varepsilon, \delta)$, on borne grossièrement le risque par $O(\tau(\varepsilon, \delta)^3)$, puis on utilise une analyse beaucoup plus fine lorsque l'algorithme est localisé autour de l'optimum, ce qui nous permet d'obtenir un terme dominant optimal:

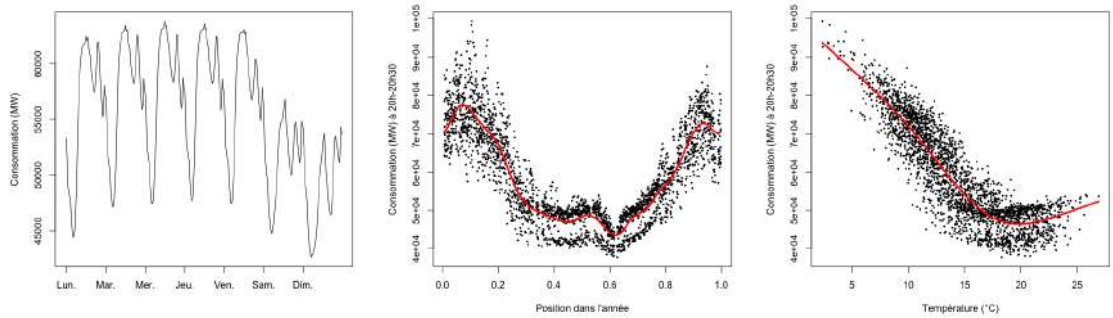


Figure 1: À gauche: profil hebdomadaire de la consommation d'électricité. On observe un comportement différent entre les jours ouvrés, le samedi et le dimanche. On se focalise ensuite à 20h-20h30 pour présenter la dépendance de la consommation d'électricité en la position dans l'année (variable linéaire allant de 0 à 1 du 1er Janvier au 31 Décembre) au centre, et en la température à droite.

Théorème 3. *Si les hypothèses 1, 2, 3 sont vérifiées, alors pour tout* $\varepsilon, \delta > 0$*, on a simultanément pour tout* $n \geq 1$*,*

$$\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^\star) \leq Cd\left(\sigma_{\text{sg}}^2 + D_{\text{app}}^2 + \varepsilon^2 D_x^2\right) \ln n + O(\ln \delta^{-1}) + O(\tau(\varepsilon, \delta)^3)\,,$$

avec probabilité au moins $1 - \delta$*, pour une constante universelle* C*.*

Il est à noter que l'algorithme ne dépend pas de ε et qu'on peut le choisir de sorte à optimiser la borne, donc décroissant en n.

Cette étude du risque cumulé est motivée par le passage au cas dynamique. En effet, si l'on suppose les données non stationnaires, il faudrait se comparer à un $\theta_t^\star$ non constant. Pour imposer une certaine régularité sur ce paramètre optimal, il serait raisonnable de supposer que le modèle espace-état est bien spécifié, et on comparerait $\hat{\theta}_t$ à θ_t. En ce sens, dans le cas statique, on s'est comparé à $\theta^\star$ minimisant le risque.

3 Prévision de consommation électrique en 2020

On s'intéresse à la consommation électrique de la France à un pas demi-horaire pendant la crise de 2020, voir Obst et al. (2020). On affiche Figure 1 la dépendance de la consommation à des variables calendaires ainsi qu'à la température.

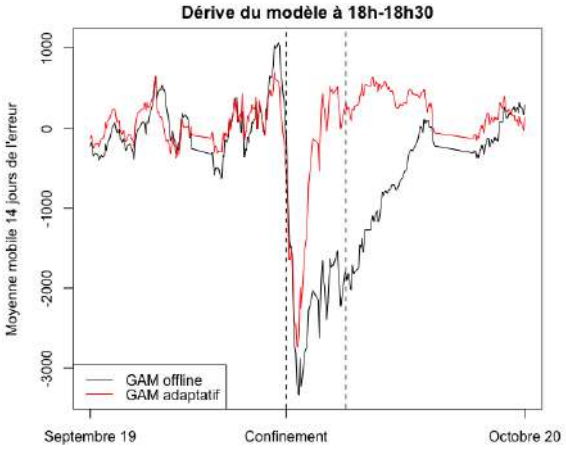


Figure 2: Évolution de l'erreur du modèle à 18h-18h30. On compare le GAM entraîné jusqu'en Septembre 2019 (GAM offline) avec une version adaptée par filtre de Kalman (Section 3.2).

3.1 Modèle additif généralisé

On utilise un modèle additif généralisé (GAM), voir Pierrot et al. (2011), très utilisé pour prévoir la consommation électrique chez EDF. Il consiste à prévoir la consommation comme une somme de fonctions non nécessairement linéaires des différentes variables explicatives. Formellement, on écrit

$$\hat{y}_t = \sum_{i=1}^{d} f_i(x_t^{(i)}) \,,$$

où les fonctions f_i sont décomposées sur des bases de spline. Les différents paramètres du modèle sont les variables calendaires, la température (ainsi que des lissages), et les valeurs précédentes de la consommation électrique.

Cependant, ce modèle est peu performant au Printemps 2020 à cause du confinement en place en France. En particulier, la consommation a chuté brusquement ce qui a entraîné une sur-prévision importante des modèles non adaptatifs, voir Figure 2. On observe en effet que le modèle additif non adaptatif dérive de façon très brutale au moment du confinement, puis la consommation semble se normaliser progressivement au cours de l'été. On se propose d'adapter le modèle au cours du temps, et le modèle obtenu subit certes une rupture au moment du confinement mais corrige le biais en quelques semaines.

3.2 Filtre de Kalman

Pour faire évoluer le modèle au cours du temps, il est naturel de le corriger à l'aide d'une combinaison linéaire en les f_i comme dans Ba. et al. (2012). Cependant plutôt que d'utiliser une pondération exponentielle pour accorder plus d'importance aux données les plus récentes, on utilise le modèle espace état suivant:

$$y_t = \theta_t^\top f(x_t) + \varepsilon_t,$$
$$\theta_{t+1} = \theta_t + \eta_t.$$

Les formules du filtre de Kalman nous permettent d'estimer récursivement l'état. Il n'existe pas de méthode idéale pour optimiser les paramètres du filtre de Kalman: les variance/covariance σ^2, Q des bruits, ainsi que le prior $\hat{\theta}_1, P_1$. En effet, la log-vraisemblance et la log-vraisemblance complète ont des optima locaux non globaux. On cherche donc des heuristiques qui permettent d'obtenir de bonnes valeurs des hyper-paramètres, sans garantie d'optimalité. En ce sens, on a essayé d'appliquer l'algorithme Expectation-Maximization, un algorithme itératif qui permet d'obtenir un maximum local de la log-vraisemblance complète. Cependant on a obtenu de meilleurs résultats avec une grid search sur Q diagonale, dans laquelle on a procédé de façon itérative en choisissant à chaque étape le coefficient qui améliorait le plus la vraisemblance.

On présentera les résultats obtenus par différents choix des hyper-paramètres, et l'on comparera avec les performances obtenus par d'autres modèles adaptatifs.

Bibliographie

Ba, A., Sinn, M., Goude, Y. and Pompey, P. (2012). Adaptive learning of smoothing functions: Application to electricity load forecasting, *Advances in neural information processing systems* (pp. 2510-2518).

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*, Springer.

Cesa-Bianchi, N. and Lugosi, G. (2006), *Prediction, Learning, and Games*, Cambridge university press.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, Oxford university press.

Obst, D., de Vilmarest, J., and Goude, Y. (2020). Adaptive Methods for Short-Term Electricity Load Forecasting During COVID-19 Lockdown in France. *arXiv preprint* (2009.06527).

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models, *Proceedings of ISAP power*.

de Vilmarest, J. and Wintenberger, O. (2020). Stochastic Online Optimization using Kalman Recursion, *arXiv preprint* (2002.03636).

Wilks' theorem for semiparametric regressions with weakly dependent data

Marie du Roy de Chaumaray [1] & Matthieu Marbac [2] & Valentin Patilea [3]

[1] *marie.du-roy-de-chaumaray@ensai.fr*
[2] *matthieu.marbac-lourdelle@ensai.fr*
[3] *valentin.patilea@ensai.fr*
[1 2 3] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*

Résumé. La vraisemblance empirique est une méthode majeure d'inférence statistique amplement étudiée lors des vingt dernières années. Ses développements rapides s'expliquent par d'importantes propriétés garanties par le fait que la vraisemblance empirique associe la flexibilité des approches non paramétriques à l'efficacité des méthodes de vraisemblance. Dans ce travail, nous considérons un modèle de régression semi-paramétrique single-index avec un processus stationnaire α-mélangeant. L'objectif est d'étudier sous quelles conditions le rapport de vraisemblance empirique converge encore vers une distribution khi-deux afin de tester la valeur des paramètres du modèle. Notons que de nombreux modèles entrent dans la classe considérée, parmi eux on peut citer les modèles de régression où les erreurs sont des différences de martingales ou des processus ARCH, les modèles CHARN (qui englobent des processus simples mais courants comme des AR).

Mots-clés. Noyau, processus faiblement dépendants, série temporelle non linéaire, inférence paramétrique, statistique pivotale,...

Abstract. The empirical likelihood inference is extended to a class of semiparametric models for stationary, weakly dependent series. A partially linear single-index regression is used for the conditional mean of the series given its past, and the present and past values of a vector of covariates. A parametric model for the conditional variance of the series is added to capture further nonlinear effects. We propose suitable moment equations which characterize the mean and variance model. We derive an empirical log-likelihood ratio which includes nonparametric estimators of several functions, and we show that this ratio behaves asymptotically as if the functions were given.

Keywords. Kernel smoothing, $\alpha-$mixing, Nonlinear time series, Nuisance function, Parametric inference, Pivotal statistic,...

We aim modeling and doing inference for one-dimensional time series (Y_i) given a vector-valued time series (V_i) and the past values of Y_i and V_i, $i \in \mathbb{Z}$. For this purpose we propose flexible semiparametric models for conditional mean and conditional variance of Y_i. Formally, let (Z_i) be a strictly stationary and strongly mixing sequence of random vectors with $Z_i = (V_i^\top, \varepsilon_i)^\top \in \mathbb{R}^{d_X+d_W} \times \mathbb{R}$ where $V_i = (X_i^\top, W_i^\top)^\top \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_W}$. Let

$(\mathcal{F}_i)$ be its natural filtration. For any positive integer r, we denote the r lagged values of Z_i by $Z_i^{\{r\}} = (V_{i-1}^\top, Y_{i-1}, \ldots, V_{i-r}^\top, Y_{i-r})^\top$.

Let us consider the semiparametric model defined by

$$Y_i = \mu(V_i; \gamma, m) + \varepsilon_i \quad \text{with} \quad \mu(V_i; \gamma, m) = l(X_i; \gamma_1) + m(W_i^\top \gamma_2), \tag{1}$$

where

$$\mathbb{E}[\varepsilon_i \mid V_i, \mathcal{F}_{i-1}] = 0, \tag{2}$$

and

$$\mathbb{E}[\varepsilon_i^2 \mid V_i, \mathcal{F}_{i-1}] = \sigma^2(V_i, Z_i^{\{r\}}; \beta), \tag{3}$$

$\gamma = (\gamma_1^\top, \gamma_2^\top)^\top$, $\theta = (\gamma^\top, \beta^\top)^\top$ and $m(\cdot)$ is an infinite dimensional parameter. Thus θ gathers the finite dimensional parameters, and our interest will focus on this vector, while $m(\cdot)$ is considered as a nuisance parameter. The value of r, as well as the real-valued functions $l(\cdot)$ and $\sigma^2(\cdot)$, are given. Moreover, the functions we consider for $\sigma^2(\cdot)$ do not require to know the infinite dimensional parameter $m(\cdot)$. Let θ_0 and $m_0(\cdot)$ denote the true values of the finite and infinite-dimensional parameters of the model, respectively. The vector V_i may include common random variables and/or lagged values of Y_i, as well as exogenous covariates. We call a model defined by (1)-(3) a CHPLSIM which stands for *Conditional Heteroscedastic Partially Linear Single-Index Model*. The methodology we will propose in the sequel allows us to replace (3) by a higher order moment equation, or to add higher order moments to (3).

CHPLSIM is related to the model proposed by [14] in the case of independent observations following the same distribution. Our model covers a wide class of models for weakly dependent and independent data. First, with $l(X_i; \gamma_1) = X_i^\top \gamma_1$, CHPLSIM includes the partially linear single-index model (PLSIM) [2] in which the errors ε_i are independent and identically distributed (i.i.d.) variables and V_i are independent covariates. Such semiparametric models were originally used to overcome the curse of dimensionality inherent to nonparametric regression on W_i by making use of a single-index $W_i^\top \gamma_2$. The PLSIM includes the partially linear models with a single variable in the nonparametric part. Our non-i.i.d. framework allows for heteroscedasticity in the errors of PLSIM, with the conditional variance of the errors possibly depending of both the covariates and the lagged errors values. For instance, it allows martingale difference errors, as considered by [5] and [7]. [24] considered a model defined by (1) for strongly mixing stationary time series, with identity function $l(\cdot)$, $X_i = W_i$ and W_i admitting a density. Their study focuses on the estimation of the parameters in the conditional mean function using kernel smoothing, without investigating the conditional variance, as allows condition (3). In the same type of model, using local linear smoothing, [23] allowed for X_i not necessarily equal to W_i and, at the price of a trimming, relaxed the condition of a density for W_i to a density for the index $W_i^\top \gamma_2$. More recently, using orthogonal series expansions, [6] extended the model defined by (1) to the case where $X_i = W_i$ is a multi-dimensional integrated process.

Model (1)-(2) is also related to and extends a large class of location-scale type models called conditionnal heteroscedastic autoregressive nonlinear (CHARN) models [9, 10]. CHARN models include many well-known models widely used with application areas as different as foreign exchange rates [1] or brain and muscular wave analysis [11]. For general nonlinear autoregressive processes, we refer to the book of [20] for the basic definitions as well as numerous applications on real data sets. More generally, nonparametric techniques for nonlinear AR processes can be found in the review of [8]. CHPLSIM allows for a semiparametric specification of the conditional mean and for exogenous covariates.

We are interested in inference on the finite dimensional parameter θ constituted of finite-dimensional parameters from both the conditional mean and the conditional variance functions. When the interest focuses on the parameters of the conditional mean, it suffices to consider equations (1)-(2) with a fully nonparametric conditional variance $\sigma^2(\cdot)$. However, in the time series context, modeling the variance can be important, for instance for forecasting purposes. For our inference purpose, we propose a semiparametric empirical likelihood approach with infinite-dimensional nuisance parameters. Empirical likelihood (EL), introduced by [17, 18], is a general inference approach for models specified by moment conditions. Under the assumption of independence between observations, empirical likelihood has been used for inference on finite dimensional parameters into regression models and unconditional moment equations. See [19]; see also the review of [4].

Under i.i.d. data assumption, [21, 22] and [15] study the conditions implying that the empirical likelihood log-ratio (ELR) still converges to a chi-squared distribution for the partially linear model. Due to the curse of the dimensionality, the performances of the nonparametric estimators decrease dramatically with the number of variables. [25] and [28] show that, if the density of the index is bounded away from zero, the ELR converges to a chi-squared distribution and thus permits parameter testing, for single-index model and PLSIM respectively (see also [27]).

The aim of this paper is to propose a novel general semiparametric regression framework for EL inference which allows for dependent data. Some related cases have been considered in the literature. For instance, the ELR with longitudinal data has been considered by [26], for the partially linear model, and by [13], for PLSIM. In their framework, the convergence of the ELR is guaranteed by the independence between individuals for which a finite bounded number of repeated observations are available. Empirical likelihood has also been used for specific models in times series (see the review of [16]; see also [3]). Most of the methods developed in this context are based on a blockwise version of empirical likelihood, first introduced by [12]. A large amount of generalizations have been proposed in the literature depending on the type of dependency. We refer to [16] for an overview of those techniques of blocking. However, in such an approach, one has to tune additional parameters such as the number, the length or the overlapping of the blocks, which might be a complex task.

Our contribution is the extension of the EL inference approach to the case of CH-

PLSIM defined by (1)-(3), for weakly dependent data. This extension is realized without imposing the density of the index bounded away from zero, as it is usually assumed in the literature in the case of i.i.d. data. See, for instance, [28], [27] and [14]. Such a very convenient, though quite stringent, condition implies a bounded support for the index, a restriction which makes practically no sense in a general time series framework. To obtain our results, a preliminary crucial step before using EL consists in building a fixed number of suitable unconditional moment equations equivalent to conditional moment equations defining the regression model. By the definition of these unconditional moment equations, our approach will not require a blocking data technique. Then, we follow the lines of [19], with the difference of the presence of infinite-dimensional nuisance parameters. We show that the nonparametric estimation of the nuisance parameters does not affect the asymptotics and the ELR still converges to a chi-squared distribution. The negligibility of the nonparametric estimation effect is obtained under mild conditions on the smoothing parameter. [3] studied the EL inference for unconditional moment equations under strongly mixing conditions, with the number of moment equations allowed to increase with the sample size. Since conditional moment equations models could be approximated by models defined by a large number of unconditional moment equations, in principle, [3] could also consider semiparametric models. However, the practical effectiveness of their approach remains an uninvestigated issue.

Our work is organised as follows. First we consider the profiling approach for the nuisance parameter $m(\cdot)$ and the identification issue for the finite-dimensional parameters. Next, we establish the equivalence between our model equations and suitable unconditional moment estimating equations. The number of unconditional equations is given by the dimension of the vector of identifiable parameters in the (CH)PLSIM. Then we establish Wilks' Theorem in our context. Our methodology is illustrated by numerical experiments and an application using daily pollution data inspired by the study of [14]. All details are available on ArXiv: `https://arxiv.org/abs/2006.06350`

References

[1] P. Bossaerts, C. Hafner, and W. Härdle. *A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series*, pages 71–83. Physica-Verlag HD, Heidelberg, 1996.

[2] R. J. Carroll, Jianqing Fan, Irène Gijbels, and M. P. Wand. Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, 92(438):477–489, 1997.

[3] Jinyuan Chang, Song Xi Chen, and Xiaohong Chen. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1):283 – 304, 2015.

[4] Song Xi Chen and Ingrid Van Keilegom. A review on empirical likelihood methods for regression. *TEST*, 18(3):415–447, 2009.

[5] Xia Chen and Hengjian Cui. Empirical likelihood inference for partial linear models under martingale difference sequence. *Statistics & Probability Letters*, 78(17):2895–2901, 2008.

[6] Chaohua Dong, Jiti Gao, and Dag TjÃ¸stheim. Estimation for single-index and partially linear single-index integrated models. *Ann. Statist.*, 44(1):425–453, 2016.

[7] Guo-Liang Fan and Han-Ying Liang. Empirical likelihood inference for semiparametric model with linear process errors. *Journal of the Korean Statistical Society*, 39(1):55–65, 2010.

[8] Wolfgang Härdle, Helmut Lütkepohl, Rong Chen, Wolfgang Hardle, and Helmut Lutkepohl. A review of nonparametric time series analysis. *International Statistical Review / Revue Internationale de Statistique*, 65(1):49–72, apr 1997.

[9] Wolfgang Härdle, Alexandre Tsybakov, and Lijian Yang. Nonparametric vector autoregression. *Journal of Statistical Planning and Inference*, 68(2):221–245, 1998.

[10] Hiroomi Kanai, Hiroaki Ogata, and Masanobu Taniguchi. Estimating function approach for charn models. *Metron*, 68(1):1–21, 2010.

[11] H. Kato, M. Taniguchi, and M. Honda. Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Transactions on Signal Processing*, 54(9):3414–3425, sep 2006.

[12] Yuichi Kitamura. Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, 25(5):2084–2102, 1997.

[13] Gaorong Li, Lixing Zhu, Liugen Xue, and Sanying Feng. Empirical likelihood inference in partially linear single-index models for longitudinal data. *Journal of Multivariate Analysis*, 101(3):718–732, 2010.

[14] Heng Lian, Hua Liang, and Raymond J. Carroll. Variance function partially linear single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):171–194, 2015.

[15] Xuewen Lu. Empirical likelihood for heteroscedastic partially linear models. *Journal of Multivariate Analysis*, 100(3):387–396, 2009.

[16] Daniel J Nordman and Soumendra N Lahiri. A review of empirical likelihood methods for time series. *Journal of Statistical Planning and Inference*, 155:1–18, 2014.

[17] Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[18] Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

[19] Jing Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325, 1994.

[20] Howell Tong. *Nonlinear time series*, volume 6 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1990. A dynamical system approach, With an appendix by K. S. Chan, Oxford Science Publications.

[21] Qi-Hua Wang and Bing-Yi Jing. Empirical likelihood for partial linear models with fixed designs. *Statistics & Probability Letters*, 41(4):425–433, 1999.

[22] Qi-Hua Wang and Bing-Yi Jing. Empirical likelihood for partial linear models. *Annals of the Institute of Statistical Mathematics*, 55(3):585–595, 2003.

[23] Yingcun Xia and Wolfgang HÃ¤rdle. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162 – 1184, 2006.

[24] Yingcun Xia, Howell Tong, and W. K. Li. On extended partially linear single-index models. *Biometrika*, 86(4):831–842, 1999.

[25] Liu-Gen Xue and Lixing Zhu. Empirical likelihood for single-index models. *Journal of Multivariate Analysis*, 97(6):1295–1312, 2006.

[26] Liugen Xue and Lixing Zhu. Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, 94(4):921–937, 2007.

[27] Lixing Zhu, Lu Lin, Xia Cui, and Gaorong Li. Bias-corrected empirical likelihood in a multi-link semiparametric model. *Journal of Multivariate Analysis*, 101(4):850 – 868, 2010.

[28] Lixing Zhu and Liugen Xue. Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):549–570, 2006.

Estimation du maximum de vraisemblance et tests d'hypothèse pour des panels de processus semi-Markoviens

Cindy Frascolla & Hervé Cardot
Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Université de Bourgogne, 21000 Dijon, France

Résumé. Nous nous intéressons à des tests d'hypothèse pour des panels de processus semi-Markoviens, motivés par une application en analyse sensorielle. Pour modéliser les différentes sensations perçues au cours de la dégustation d'un produit, Lecuelle *et al.* (2018) ont proposé d'utiliser les processus semi-Markoviens. Pour déterminer si deux produits testés sont perçus différemment, un test statistique basé sur le rapport de vraisemblance a été construit et étudié par simulations (Frascolla *et al.* 2020). Nous étudions dans ce travail la convergence asymptotique lorsque le nombre L de trajectoires tends vers l'infini des estimateurs du maximum de vraisemblance des paramètres des processus semi-Markoviens et la distribution asymptotique du rapport de vraisemblance. Nous considérons deux modèles d'observation pour chaque trajectoire : celui d'un processus semi-Markovien absorbant et celui où chaque trajectoire est composée d'un nombre aléatoire de transitions.

Mots-clés. Analyse sensorielle, Estimateur du maximum de vraisemblance, Processus semi-Markoviens, Statistique asymptotique, Test du rapport de vraisemblance

Abstract. We are interested in hypothesis testing for panels of semi-Markov processes motivated by an application in sensory analysis. To model the different sensations perceived during the tasting of a product, Lecuelle *et al.* (2018) have considered semi-Markov processes. To determine if two tested products are felt differently a statistical test based on the likelihood ratio has been built and studied by simulations (Frascolla *et al.* 2020). The aim of this work is to study the asymptotic convergence of the maximum likelihood estimators of the parameters of the semi-Markov processes and the asymptotic distribution of the likelihood ratio when the number L of trajectories tends to infinity. We consider two observation designs for each trajectory: one with an absorbing semi-Markov process and one where each trajectory is composed of a random number of transitions.

Keywords. Sensory analysis, Maximum likelihood estimator, Semi-Markov processes, Asymptotic statistics, Likelihood ratio test

1 Introduction

Ce travail est motivé par une application en analyse sensorielle dont l'objectif est de mieux comprendre les préférences des consommateurs. Lors d'une étude d'analyse sensorielle,

différents produits d'une même catégorie sont testés et les sujets indiquent la séquence des sensations perçues au cours du temps parmi une liste de descripteurs. Pour prendre en compte la dynamique du modèle (les transitions d'un descripteur vers un autre) et les temps de séjour associés, Lecuelle *et al.* (2018) ont proposé de modéliser ces données par des processus semi-Markoviens.

Les processus semi-Markoviens sont une généralisation des chaînes de Markov et permettent de considérer des modèles plus flexibles pour la loi des temps de séjour. Les livres de Limnios et Oprişan (2001) et de Barbu et Limnios (2008) présentent la théorie des processus semi-Markoviens et leur application en fiabilité et analyse de l'ADN.

Une des principales questions en analyse sensorielle est de déterminer si deux produits testés sont différents. Pour répondre à cette question, un test statistique basé sur le rapport de vraisemblance a été proposé dans Frascolla *et al.* (2020) avec trois approches pour déterminer la zone de rejet : deux approches basées sur des techniques de ré-échantillonnage (bootstrap paramétrique et permutations) et une approche asymptotique basée sur la loi du rapport de vraisemblance en supposant de manière intuitive qu'elle suivait une loi du χ^2, ce qui a été vérifié sur des simulations. L'objectif de ce travail est de montrer la consistance des estimateurs de vraisemblance et leur normalité asymptotique puis d'étudier la loi asymptotique du rapport de vraisemblance.

2 Modèle et notations

2.1 Définition des processus semi-Markoviens

Soit $Z = (Z_t)_{t\in\mathbb{R}_+}$ un processus stochastique à valeur dans $E = \{1, \ldots, D\}$ avec $D < +\infty$. Soient $J = (J_n)_{n\in\mathbb{N}}$ la suite des états visités par Z et $X = (X_n)_{n\in\mathbb{N}^*}$ la suite des temps de séjour associés.

Nous supposons que le processus $(J_n, X_n)_{n\geq 1}$ vérifie la propriété de Markov, pour $t \in T = [0, +\infty[$, $\ell \in E$ et $j \neq \ell$,

$$\mathbb{P}(J_{n+1} = j, X_{n+1} \leq t \mid J_0, \ldots, J_n, X_1, \ldots, X_n) = \mathbb{P}(J_{n+1} = j, X_{n+1} \leq t \mid J_n).$$

Le processus $(J_n, X_n)_{n\geq 1}$ est un processus de renouvellement de Markov tandis que le processus donnant l'état visité à chaque instant t est appelé processus semi-Markovien (voir par exemple Limnios et Oprişan, 2001).

La loi du processus semi-Markovien est caractérisée par sa distribution initiale $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$ avec $\alpha_j = \mathbb{P}(J_0 = j)$, $j = 1, \ldots, D$ et par son noyau semi-Markovien

$$Q_{ij}(t) = \mathbb{P}(J_n = j, X_n \leq t \mid J_{n-1} = i)$$

avec la convention $X_0 = S_0 = 0$.

La matrice de transition de la chaîne de Markov $(J_n)_{n\geq 1}$ est notée $\mathbf{P}$ et est constituée des éléments $p_{ij} = \mathbb{P}(J_n = j \mid J_{n-1} = i)$, pour $i \neq j \in E \times E$ et $p_{ii} = 0$ pour tout $i \in E$.

Soit $W_{ij}(t)$ la fonction de répartition des temps de séjour vérifiant :

$$W_{ij}(t) = \mathbb{P}(X_n \leq t \mid J_{n-1} = i, J_n = j), \quad t \geq 0, i \neq j.$$

Par la propriété de Markov on a $Q_{ij}(t) = p_{ij}W_{ij}(t)$. On suppose que les distributions des temps de séjour appartiennent à une famille paramétrique de densité. Pour $i \neq j \in E \times E$, on note $f(t; \theta_{ij})$ la densité du temps de séjour de $X_n | J_{n-1} = i, J_n = j$ avec $\theta_{ij} \in \mathbb{R}^d$. On note $\boldsymbol{\theta} = (\theta_{ij}, i \neq j \in E \times E)$.

Ainsi, la distribution du processus semi-Markovien Z est caractérisée par le vecteur des paramètres $(\boldsymbol{\alpha}_0, \mathbf{P}_0, \boldsymbol{\theta}_0)$.

2.2 Les différents modèles d'observation considérés

On considère des panels constitués de L trajectoires indépendantes, $S_1, \ldots, S_L$, issues du même processus semi-Markovien de paramètre $(\boldsymbol{\alpha}_0, \mathbf{P}_0, \boldsymbol{\theta}_0)$ où une séquence S est constituée des différents états visités par le processus et des temps de séjour associés.

Deux modèles d'observation sont considérés : un premier modèle où chaque séquence est constituée d'un nombre aléatoire de transitions et un second modèle où l'espace d'état contient un état absorbant et chaque séquence s'arrête une fois qu'elle a atteint cet état absorbant. Dans le premier cas, on note M le nombre aléatoire de transitions et on suppose que M est strictement positif et d'espérance finie. Dans le deuxième cas on suppose que l'état absorbant est accessible et on ré-ordonne les états de E de sorte que l'état absorbant soit le dernier état, c'est-à-dire l'état D. On suppose de plus que le processus débute dans un état non absorbant, c'est-à-dire $\alpha_D = \mathbb{P}(J_0 = D) = 0$, et on note τ le nombre de transitions jusqu'à absorption. La variance et l'espérance de τ pour des chaînes de Markov absorbantes sont finies (voir Kemeny et Snell (1976) par exemple).

Le choix d'étudier ces deux différents modèles repose sur leur application en analyse sensorielle. En effet, pour certaines études d'analyse sensorielle, l'étude prend fin une fois que le sujet clique sur un bouton "STOP" qui peut être considéré comme un état absorbant. Pour d'autres études, elle prend fin quand le sujet ne ressent plus rien dans ce cas on a un nombre aléatoire de transitions.

Dans le cas d'un processus semi-Markovien avec un nombre aléatoire M de transitions, la vraisemblance d'une séquence $S = \{j_0, x_1, j_1, \ldots, j_{M-1}, x_M, j_M\}$ s'écrit : $\mathcal{L}(S; \boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta}) = \alpha_{j_0} \prod_{k=1}^{M} p_{j_{k-1}j_k} f\left(x_k; \theta_{j_{k-1}j_k}\right)$. Le processus de comptage $N_{ij}^{(\ell)}$ donnant le nombre de transition de i vers j pour une séquence ℓ s'écrit $N_{ij}^{(\ell)} = \sum_{n=0}^{m_\ell - 1} \mathbf{1}_{\{J_n^{(\ell)} = i, J_{n+1}^{(\ell)} = j\}}$ et celui donnant le nombre de visite de l'état i pour une séquence ℓ noté $N_i^{(\ell)}$ s'écrit $N_i^{(\ell)} = \sum_{n=0}^{m_\ell - 1} \mathbf{1}_{\{J_n^{(\ell)} = i\}}$ avec m_ℓ le nombre de transitions observées pour la trajectoire S_ℓ.

Dans le cas d'un processus semi-Markovien absorbant, la vraisemblance de la séquence séquence $S = \{j_0, x_1, j_1, \ldots, j_{\tau-1}, x_\tau, j_\tau\}$ est $\mathcal{L}(S; \boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta}) = \alpha_{j_0} \prod_{i=1}^{\tau} p_{j_{i-1}j_i} f\left(x_i; \theta_{j_{i-1}j_i}\right)$. Les processus de comptage $N_{ij}^{(\ell)}$ et $N_i^{(\ell)}$ s'écrivent alors $N_{ij}^{(\ell)} = \sum_{n=0}^{\tau_\ell - 1} \mathbf{1}_{\{J_n^{(\ell)} = i, J_{n+1}^{(\ell)} = j\}}$ et $N_i^{(\ell)} = \sum_{n=0}^{\tau_\ell - 1} \mathbf{1}_{\{J_n^{(\ell)} = i\}}$.

3 Convergence asymptotique de l'estimateur du maximum de vraisemblance

La vraisemblance des L séquences s'écrit $\mathcal{L}(S_1,\ldots,S_L;\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta}) = \prod_{l=1}^{L}\mathcal{L}(S_l;\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta})$ où $\mathcal{L}(S_\ell;\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta})$ est la vraisemblance de la séquence ℓ. Soit $\hat{Q}_L(\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta})$ la valeur moyenne de la log-vraisemblance, sur les trajectoires $S_1,\ldots,S_L$, définie par :

$$\begin{aligned}\widehat{Q}(\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta}) &= \frac{1}{L}\sum_{\ell=1}^{L}\ln\mathcal{L}(S_\ell;\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta}) \\ &= \frac{1}{L}\sum_{\ell=1}^{L}\ln\left(\alpha_{j_0^{(\ell)}}\right) + \widehat{Q}_{\mathbf{P}}(\mathbf{P}) + \widehat{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \qquad (1)\end{aligned}$$

avec

$$\widehat{Q}_{\mathbf{P}}(\mathbf{P}) = \frac{1}{L}\sum_{\ell=1}^{L}\left\{\sum_{i\in E}\sum_{\substack{j=1\\ j\neq i}}^{D-1}\left[N_{ij}^{(\ell)}\ln(p_{ij})\right] + \sum_{i\in E}\left(N_i^{(\ell)} - \sum_{k=1}^{D-1}N_{ik}^{(\ell)}\right)\ln\left(1-\sum_{j=1}^{D-1}p_{ij}\right)\right\}$$

et

$$\widehat{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{\substack{i,j\in E\\ j\neq i}}\Big[\sum_{k\in\{N_{ij}^{(\ell)}\}}\ln\left(f(x_{ij}^{(\ell,k)};\theta_{ij})\right)\Big].$$

où $\{N_{ij}^{(\ell)}\} = \{1,\ldots,N_{ij}^{(\ell)}\}$ si $N_{ij}^{(\ell)} \geq 1$ et $\{N_{ij}^{(\ell)}\} = \emptyset$ sinon et $x_{ij}^{(\ell,k)}$ le temps de séjour dans l'état i avant d'aller dans l'état j pendant la visite numéro k.

On remarque à partir de (1) que la maximisation de la log-vraisemblance en $(\boldsymbol{\alpha},\mathbf{P},\boldsymbol{\theta})$ s'effectue de manière indépendante pour chaque ensemble de paramètres. On note $\widehat{\boldsymbol{\alpha}}$, $\widehat{\mathbf{P}}$ et $\widehat{\boldsymbol{\theta}}$ les estimateurs de $\boldsymbol{\alpha}$, $\mathbf{P}$ et $\boldsymbol{\theta}$.

L'étude des propriétés asymptotiques de l'estimateur du maximum de vraisemblance peut donc être effectuée en trois parties en analysant séparément chaque ensemble de paramètres. L'estimateur $\widehat{\boldsymbol{\alpha}}$ est simplement l'estimateur du maximum de vraisemblance pour une distribution multinomiale (voir Trevezas et Limnios 2011 par exemple).

La difficulté liée à l'estimation des paramètres $\mathbf{P}$ et $\boldsymbol{\theta}$ provient du fait que les sommations s'effectuent sur un nombre aléatoire d'indices, qui n'est pas nécessairement indépendant des variables aléatoires étudiées. Il faut également que les conditions classiques assurant la convergence des estimateurs du maximum de vraisemblance soient vérifiées : conditions sur la loi des temps de séjour (voir les Lemmes 2.2, 2.4 et le Théorème 2.1 Newey et McFadden (1994)) et condition que les p_{ij} soient compris dans l'intervalle $]0,1[$.

Pour les deux modèles d'observation la consistance de $\widehat{\mathbf{P}}$ découle de la loi forte des grands nombres, du *"continuous mapping theorem"* et de l'identité de Wald (voir Gut

2009, Chapitre 1). Concernant l'estimation du paramètre $\boldsymbol{\theta}$, l'identité de Wald et les théorèmes de Newey et McFadden (1994) permettent de démontrer la convergence en probabilité de $\widehat{\boldsymbol{\theta}}$ vers $\boldsymbol{\theta}_0$. Le théorème central limite et le théorème d'Anscombe pour le cas multidimensionnel (voir Gut 2009) permettent d'obtenir la normalité asymptotique de $\hat{\mathbf{P}}$. La normalité asymptotique de $\hat{\boldsymbol{\theta}}$ est obtenue grâce au Théorème 3.1 de Newey et McFadden (1994) appliqué à $\widehat{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

4 Test du rapport de vraisemblance pour comparer deux populations

On suppose maintenant que les L trajectoires proviennent de deux populations, et on dispose de deux échantillons de taille L_1 et L_2 issus de deux processus semi-Markoviens Z^1 et Z^2 caractérisés par les vecteurs de paramètres $(\boldsymbol{\alpha}_1, \mathbf{P}_1, \boldsymbol{\theta}_1)$ et $(\boldsymbol{\alpha}_2, \mathbf{P}_2, \boldsymbol{\theta}_2)$. On souhaite tester l'égalité des lois de ces deux processus et on considère les hypothèses :

$$\begin{aligned} H_0 &: (\boldsymbol{\alpha}_1, \mathbf{P}_1, \boldsymbol{\theta}_1) = (\boldsymbol{\alpha}_2, \mathbf{P}_2, \boldsymbol{\theta}_2) \\ H_1 &: (\boldsymbol{\alpha}_1, \mathbf{P}_1, \boldsymbol{\theta}_1) \neq (\boldsymbol{\alpha}_2, \mathbf{P}_2, \boldsymbol{\theta}_2). \end{aligned}$$

La statistique de test basée sur le rapport de vraisemblance, notée LR s'écrit :

$$LR = \frac{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\boldsymbol{\alpha}}, \hat{\mathbf{P}}, \hat{\boldsymbol{\theta}}) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\boldsymbol{\alpha}}, \hat{\mathbf{P}}, \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\boldsymbol{\alpha}}_1, \hat{\mathbf{P}}_1, \hat{\boldsymbol{\theta}}_1) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\boldsymbol{\alpha}}_2, \hat{\mathbf{P}}_2, \hat{\boldsymbol{\theta}}_2)}$$

avec $(\hat{\boldsymbol{\alpha}}_1, \hat{\mathbf{P}}_1, \hat{\boldsymbol{\theta}}_1)$ l'estimateur du maximum de vraisemblance sous H_1 pour le premier échantillon, $(\hat{\boldsymbol{\alpha}}_2, \hat{\mathbf{P}}_2, \hat{\boldsymbol{\theta}}_2)$ l'estimateur du maximum de vraisemblance sous H_1 pour le deuxième échantillon et $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{P}}, \hat{\boldsymbol{\theta}})$ l'estimateur du maximum de vraisemblance sous H_0 pour les deux échantillons.
Nous étudions également le cas de l'égalité partielle en se concentrant uniquement sur les probabilités de transition ou sur la distribution des temps de séjour grâce à deux tests partiels. Dans ce cas, la statistique de test est simple grâce à la structure multiplicative de la vraisemblance qui permet de restreindre le calcul du rapport de vraisemblance sur un sous-ensemble des paramètres.

Dans le cas du test partiel sur le paramètre $\mathbf{P}$, les hypothèses sont :

$$\begin{aligned} H_0 &: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2, \mathbf{P}_1 = \mathbf{P}_2, \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \\ H_1 &: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2, \mathbf{P}_1 \neq \mathbf{P}_2, \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \end{aligned}$$

Le rapport de vraisemblance devient alors,

$$LR = \frac{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\mathbf{P}}) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\mathbf{P}})}{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\mathbf{P}}_1) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\mathbf{P}}_2)}.$$

Dans le cas du test partiel sur le paramètre $\boldsymbol{\theta}$, les hypothèses sont :

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2, \mathbf{P}_1 = \mathbf{P}_2, \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$
$$H_1 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2, \mathbf{P}_1 = \mathbf{P}_2, \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2.$$

Le rapport de vraisemblance devient alors,

$$LR = \frac{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\boldsymbol{\theta}}) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{L_1} \mathcal{L}(S_i^1; \hat{\boldsymbol{\theta}}_1) \prod_{i=1}^{L_2} \mathcal{L}(S_i^2; \hat{\boldsymbol{\theta}}_2)}.$$

Dans un contexte de taille d'échantillon aléatoire, en adaptant les preuves de Ferguson (1996), on montre, sous des hypothèses classiques pour les lois des temps de séjour pour le maximum de vraisemblance, que $-2\ln(LR)$ converge en loi vers une loi du χ^2.

Bibliographie

Barbu, V. S. and Limnios, N. (2008). Semi-Markov chains and hidden semi-Markov models toward applications : their use in reliability and DNA analysis. *New York: Springer Science + Business Media.*

Ferguson, T. S. (1996). *A course in large sample theory.* Texts in Statistical Science Series. Chapman & Hall, London.

Frascolla, C., Lecuelle, G., Cardot, H., and Schlich, P. (2020). Two sample tests for semi-Markov processes: an application in sensory analysis. Article soumis, Institut de Mathématiques de Bourgogne, Université de Bourgogne

Gut, A. (2009). *Stopped random walks.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition.

Kemeny, J.G, Snell, J.L. (1976). *Finite Markov Chains.* Springer-Verlag, New York-Heidelberg.

Lecuelle, G., Visalli, M., Cardot, H. and Schlich, P. (2018). Modeling temporal dominance of sensations with semi-Markov chains. *Food Quality and Preference* 67, 59–66.

Limnios, N. et G. Oprişan (2001). *Semi-Markov Processes and Reliability.* Statistics for Industry and Technology. Birkhäuser Boston, Inc., Boston, MA.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. IV, pages 2111–2245. North-Holland, Amsterdam.

Trevezas, S. and Limnios, N. (2011). Exact MLE and asymptotic properties for nonparametric semi-Markov models. *Journal of Nonparametric Statistics* , 23, 952-958.

Introducing group-sparsity and orthogonality constraints in RGCCA

Vincent Guillemot[1], Arnaud Gloaguen[2], Arthur Tenenhaus[2], Cathy Philippe[3] and Hervé Abdi[4]

[1] *Hub de Bioinformatique et Biostatistique, Institut Pasteur, Paris, FR*
[2] *Laboratoire des Signaux et Systèmes, CentraleSupelec, Gif-Sur-Yvette, FR*
[3] *Neurospin, CEA, Gif-Sur-Yvette, FR*
[4] *The University of Texas at Dallas, Richardson, TX, USA*

Résumé. RGCCA est une méthode flexible et rapide qui—généralisant de nombreuses méthodes existantes—permet l'analyse de données structurées en plusieurs blocs hétérogènes. Nous présentons l'ajout dans RGCCA de deux nouvelles contraintes : une contrainte de parcimonie de groupes et une contrainte d'orthogonalité sur les poids de RGCCA. Ces deux contraintes ont pour but d'augmenter l'interprétabilité de l'analyse de données de grande dimension qui possèdent une structure de groupe. Nous appliquons cette nouvelle méthode—abrégée en gSGCCA—à l'analyse de données de gliome malin pédiatrique structurées en trois blocs. Nous montrons sur ces données le gain en interprétabilité apporté par les contraintes de parcimonie et d'orthogonalité.

Mots-clés. RGCCA, parcimonie, parcimonie de groupe, structure

Abstract. RGCCA—a fast and flexible method—generalizes many other well-known methods in order to analyze data-sets comprising multiple blocks of variables. Here we extend RGCCA by adding two new constraints to the RCCCA optimization problem: 1) group sparsity and 2) orthogonality of the block weight vectors. These two constraints facilitate the interpretability of the results when analyzing high dimensional data with a group structure. We illustrate this new method—called gSGCCA—with the analysis of pediatric high-grade glioma data: a set comprising three data blocks. This analysis shows that these new constraints greatly improve the interpretability of the statistical analysis.

Keywords. RGCCA, sparsity, group-sparsity, structure

1 Introduction

Regularized Generalized Canonical Correlation Analysis (RGCCA) [8, 9, 3] is a recent multiblock component method that generalizes traditional component-based two table methods—such as partial least square correlation, redundancy analysis, and canonical correlation—in order to analyze data sets comprising multiple blocks of data. Just like with other component methods, RGCCA results are often difficult to interpret when there

are (too?) many variables; to mitigate this problem, RGCCA has been extended to become Sparse General Canonical Correlation Analysis (SGCCA) [7]: a version of RGCCA that incorporate an ℓ_1-norm based constraint in order to generate sparse block weight vectors. This sparsification constraint improves the interpretation of the results (because it selects important variables) but at a cost: the block weight vectors are not orthogonal—a pattern that often makes the results difficult to interpret. This trade-off between sparsity and orthogonality is not specific to RGGCA: it affects all component based-methods, especially those based on the singular value decomposition (SVD) and its extensions (e.g., the generalized SVD, GSVD). Recently, however, we found that this trade-off could be eliminated 1) for the SVD: the constrained SVD (CSVD) [4], combines orthogonality and sparsity constraints to the plain SVD, and 2) for the GSVD (including block constraints on observations and variables): as implemented in sparse Multiple Correspondence Analysis (sMCA) [5].

Here, we propose to extend the approach used for the GSVD to create gSGCCA: the version of RGCCA that includes 1) a group sparsity constraint (and it associated group sparse projection), and 2) an orthogonality constraint on the block weight vectors. To do so, we applied the same group projection as in sparse MCA, combined with an orthogonality projection with projections onto convex sets (POCS) [1]. We illustrate gSGCCA with the analysis of the pediatric glioma data used in [7].

2 Method

Group sparse GCCA (gSGCCA) is defined as the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{\operatorname{argmax}} f\left(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J\right) = \sum_{\substack{j,k=1 \\ j\neq k}}^{J} c_{jk} g\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k\right)\right) \\ &\text{subject to } \begin{cases} \|\mathbf{a}_j\|_2 = 1 \\ \|\mathbf{a}_j\|_{\mathcal{G}_j} \leq s_j \quad , \ \forall j = 1,\ldots,J. \\ \mathbf{a}_j \perp \mathbf{A}_j \end{cases} \end{aligned} \tag{1}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_J$ are J centered blocks of data, the function g is defined as any continuously differentiable convex function, and the design matrix $\mathbf{C} = \{c_{jk}\}$ is a symmetric $J \times J$ matrix of non-negative elements describing the network of connections between blocks that are to be taken into account. Moreover, $\mathbf{a}_1,\ldots,\mathbf{a}_J$ are block weight vectors (i.e.. the weights applied to each block to obtain the block components), $\mathbf{A}_1,\ldots,\mathbf{A}_J$ are the previously estimated weight vectors combined into matrices, $\mathcal{G}_1,\ldots,\mathcal{G}_J$ are the groups of variables for a block, $s_1,\ldots,s_J$ are positive scalars controlling the group sparsity constraint for the block weight vectors, and the group norm is defined as: $\|\mathbf{x}\|_{\mathcal{G}} = \sum_{g=1}^{G} \|\mathbf{x}_{\iota_g}\|_2$, where $\mathbf{x}_{\iota_g}$ is the subvector of $\mathbf{x}$ that contains only the elements of group $\mathcal{G}_j$. The $\ell_{1,2}$-ball

associated with this norm is noted $\mathscr{B}_{1,2}(\cdot)$. In the next section, we present a monotone convergent algorithm for solving optimization Problem (1).

2.1 The gSGCCA algorithm

The maximization of function f over the parameter vectors $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_L)$, is implemented using cyclic Block Coordinate Ascent (BCA [2]); a procedure that updates in turn, each of the parameter vectors while keeping the others fixed. Specifically, let $\nabla_j f(\mathbf{a})$ be the partial gradient of $f(\mathbf{a})$ with respect to $\mathbf{a}_j$. We want to find an update $\hat{\mathbf{a}}_j \in \Omega_j = \{\|\mathbf{a}_j\|_2 = 1, \text{ and } \|\mathbf{a}_j\|_{\mathcal{G}_j} \leq s_j, \text{ and } \mathbf{a}_j \perp \mathbf{A}_j\}$ such that $f(\mathbf{a}) \leq f(\mathbf{a}_1, ..., \mathbf{a}_{j-1}, \hat{\mathbf{a}}_j, \mathbf{a}_{j+1}, ..., \mathbf{a}_J)$. Because f is a continuously differentiable multi-convex function and because a convex function lies above its linear approximation at $\mathbf{a}_j$ for any $\tilde{\mathbf{a}}_j \in \Omega_j$, the following inequality holds:

$$f(\mathbf{a}_1, ..., \mathbf{a}_{j-1}, \tilde{\mathbf{a}}_j, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_J) \geq f(\mathbf{a}) + \nabla_j f(\mathbf{a})^\top (\tilde{\mathbf{a}}_j - \mathbf{a}_j) := \ell_j(\tilde{\mathbf{a}}_j, \mathbf{a}). \tag{2}$$

On the right-hand side of (2), only the term $\nabla_j f(\mathbf{a})^\top \tilde{\mathbf{a}}_j$ is relevant to $\tilde{\mathbf{a}}_j$ and, so, the solution maximizing the minorizing function $\ell_j(\tilde{\mathbf{a}}_j, \mathbf{a})$ over $\tilde{\mathbf{a}}_j \in \Omega_j$ is obtained by considering:

$$\hat{\mathbf{a}}_j = \underset{\tilde{\mathbf{a}}_j \in \Omega_j}{\operatorname{argmax}} \nabla_j f(\mathbf{a})^\top \tilde{\mathbf{a}}_j = \underset{\tilde{\mathbf{a}}_j \in \Omega_j}{\operatorname{argmin}} \|\nabla_j f(\mathbf{a}) - \tilde{\mathbf{a}}_j\|_2^2. \tag{3}$$

This last equality follows from $\|\mathbf{a}_j\|_2 = 1$ as $\mathbf{a}_j \in \Omega_j$. This core optimization problem is a projection onto the intersection between the ball defined by the groups, the ℓ_2-ball, and the space orthogonal to the already estimated block weight vectors, assembled in $\mathbf{A}_j$. This projection on $\mathscr{B}_{1,2}(s_j) \cap \mathscr{B}_2(1) \cap \mathbf{A}_j^\perp$ is performed using POCS with two components: the projection onto the intersection of the group ball and the ℓ_2-ball, and the projection onto the orthogonal spaces defined by the already estimated loading vectors combined in the matrix $\mathbf{A}_j$. The complete gSGCCA algorithm is presented in Algorithm 1.

3 Application on glioma data

We applied gSGCCA to the glioma data previously analyzed with SGCCA ([7, 6]). This data-set comprises three blocks of variables: 1) gene expression data (GE), 2) comparative genomic hybridization data (CGH) and, 3) the location of the tumor in the brain. Here, we focused on the analysis of only six groups of genes highly associated with different types of brain tumors and with brain tumor development. The six groups are defined similarly for both the GE and CGH blocks. For this analysis, we used three different versions of RGCCA: 1) a classical three-block RGCCA with a complete design (i.e., all blocks are inter-connected); 2) a structured version of RGCCA where each group of genes is a block (here the design is complete within each type of data, GE or CGH, and all the

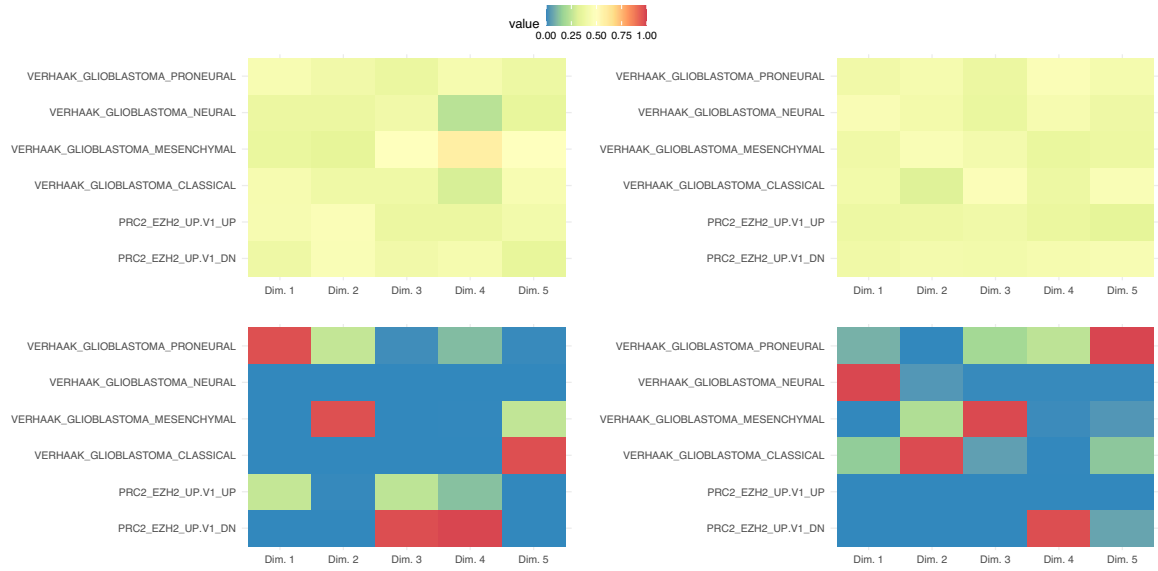


Figure 1: Comparison of the block weight vectors, summarized by groups, without sparsity constraints (upper graphs) or with a maximum sparsity constraint (lower graphs), for the GE block (on the left) and the CGH block (on the right).

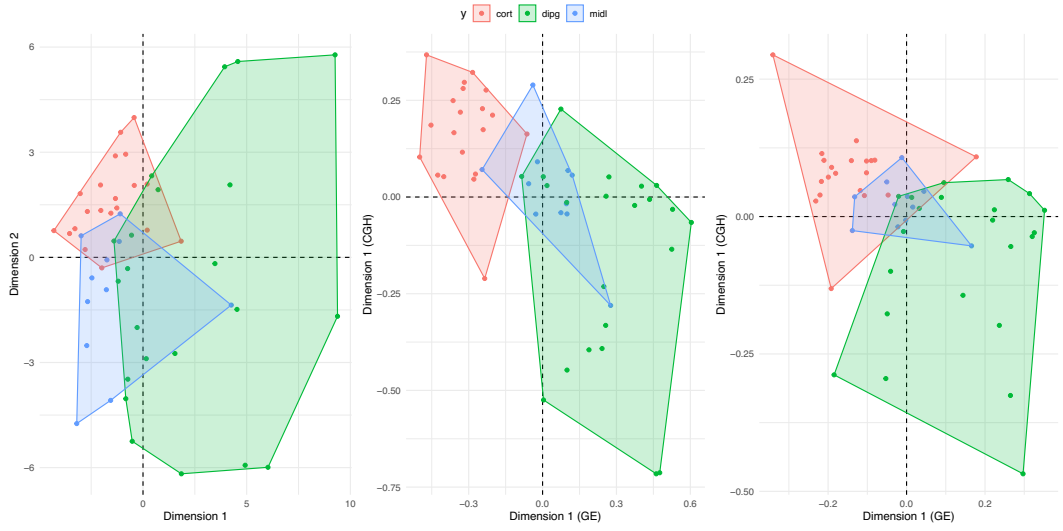


Figure 2: Factor scores of three different versions of RGCCA. Left: the two first dimensions of RGCCA applied to a 13 block dataset, 6 blocks for the GE functional groups, 6 blocks for the CGH functional groups and 1 block for the response, followed by a principal component analysis of the resulting GE and CGH block components. Middle and right: the first dimension of CGH (y-axis) as a function of the first dimension of GE (x-axis). Middle: RGCCA with no sparsity. Right: gSGCCA with maximum sparsity.

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_J$, $\mathcal{G}_1$, $\ldots$, $\mathcal{G}_J$, ε, R, $s_{1,\ell}$, $\ldots$, $s_{J,\ell}$.
Initialization: $\forall j = 1, \ldots, J, \mathbf{A}_j \leftarrow [\,]$;
Result: The estimated weight vectors combined into matrices $\mathbf{A}_1, \ldots, \mathbf{A}_J$
for $\ell = 1, \ldots, R$ **do**
  Initialize $\mathbf{a}_j^0$ for all j;
  $s \leftarrow 0$;
  while $\left\|\mathbf{a}_j^{(s+1)} - \mathbf{a}_j^{(s)}\right\| \geq \varepsilon$, $\forall j = 1, \ldots, J$ **do**
    for $j = 1, \ldots, J$ **do**
      Compute the inner component:

$$\nabla_j^s f \leftarrow \frac{1}{n}\mathbf{X}_j^t \left[\sum_{k=1}^{j-1} c_{jk} g'\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j^s,' \mathbf{X}_k\mathbf{a}_k^{s+1}\right)\right)\mathbf{X}_k\mathbf{a}_k^{s+1} \;+ \sum_{k=j+1}^{J} c_{jk} g'\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j^s, \mathbf{X}_k\mathbf{a}_k^s\right)\right)\mathbf{X}_k\mathbf{a}_k^s\right]$$

      Compute the outer weight:

$$\mathbf{a}_j^{s+1} \leftarrow \operatorname{proj}\left(\nabla_j^s f, \mathscr{B}_{1,2}(s_{j,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{A}_j^{\perp}\right)$$

    end
    $s \leftarrow s + 1$;
  end
  $\forall j = 1, \ldots, J, \mathbf{A}_j \leftarrow \left[\mathbf{A}_j, \mathbf{a}_j^{(s+1)}\right]$;
end

Algorithm 1: General algorithm of gSGCCA implementing group-sparsity and orthogonality of the block weight vectors.

GE and CGH blocks are connected to the location block), and 3) gSGCCA with maximum sparsity and a complete design (like option 1).

The block weight vectors are shown in Figure 1 for Versions 1 and 3. Each functional group is represented by its norm. This figure shows that incorporating a group-sparsity constraint in RGCCA greatly improves the interpretability of the block weight vectors because with gSGCCA only a handful of groups were selected for each dimension of GE and CGH.

The block components are shown on Figure 2, which shows that the improvement in interpretability for the loadings—observed on the block weight vectors—comes at the cost of a diminished class separation observed on the observations (this effect occurs because sparsifying the loadings automatically reduces the variance of the observations factor scores).

4 Conclusion and perspectives

We present in this paper a new method—called gSGCCA—that adds group-sparsity and orthogonality constraints to RGCCA. The application of gSGCCA to a medical example illustrates that, compared to the original RGCCA, gSGCCA provides results easier to interpret.

Future work will focus on developing a user friendly framework for the selection of the sparsity parameters to achieve some optimum trade-off between sparsity and prediction performance. We will also work on including metrics in gsGCCA to generalize its application to a wider range of data types.

References

[1] P.L. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, 1993.

[2] J. De Leeuw. Block-relaxation algorithms in statistics. In Hans-Hermann Bock, Wolfgang Lenski, and Michael M. Richter, editors, *Information Systems and Data Analysis*, pages 308–324, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg.

[3] I. Garali et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings in Bioinformatics*, 19:1356–1369, 2018.

[4] V. Guillemot et al. A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE*, 14:e0211463, 2019.

[5] V. Guillemot et al. Sparse Multiple Correspondence Analysis. In *52èmes Journées de Statistique*, Nice, France, 2020.

[6] S. Puget et al. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. *PloS one*, 7:e30313, 2012.

[7] A. Tenenhaus et al. Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*, 15:569–83, 2014.

[8] A. Tenenhaus and M. Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76:257–284, 2011.

[9] M. Tenenhaus, A. Tenenhaus, and P.J.F. Groenen. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, 82:737–777, 2017.

Spatial non-stationary modelling of extreme precipitation in the Mediterranean region

Hela Hammami [1,2] Julie Carreau [1] Luc Neppel [1] Sadok Elasmi [2]

hammamihela2@gmail.com *julie.carreau@ird.fr* *luc.neppel@umontpellier.fr* *elasmi@supcom.tn*

[1] *HydroSciences Montpellier, U. of Montpellier, CNRS, IRD, Montpellier, France.*
[2] *COSIM Lab, Higher School of Communication of Tunis, U. of Carthage, Tunis, Tunisia.*

Résumé. La région méditerranéenne est victime de plusieurs catastrophes naturelles telles que les pluies intenses qui peuvent déclencher des inondations dévastatrices. Pour cette raison, des scénarios à haute résolution spatio-temporelle sont nécessaires pour mettre en œuvre des outils d'aide à la décision pour une meilleure gestion des ressources et des risques. La théorie des valeurs extrêmes (EVT) est une branche de la statistique qui fournit un cadre approprié pour la modélisation de tels événements. Par ailleurs, la distribution des précipitations est non-stationnaire dans l'espace vue l'hétérogénéité spatiale de la zone méditerranéenne. Pour prendre en compte la non-stationnarité spatiale, les paramètres de la distribution peuvent être considérés comme des fonctions de covariables. Ces fonctions peuvent être mises en oeuvre avec des modèles non linéaires non paramétriques flexibles tels que les réseaux de neurones artificiels (ANN). L'enjeu principal dans cette analyse est la sélection du niveau de complexité de l'ANN (i.e. le nombre de neurones cachés) et le choix optimal des covariables fournis en entrée de l'ANN. Trois sites sont considérés formant un gradient d'aridité nord-sud. Les résultats présentés concernent principalement le site le plus au sud en Tunisie centrale.

Mots-clés. Événements de précipitations intenses, non-stationnarité spatiale, théorie des valeurs extrêmes, réseaux de neurones artificiels.

Abstract. The Mediterranean region is a victim of several natural disasters such as heavy rains which can trigger devastating floods. For this reason, scenarios with high spatiotemporal resolution are required to implement decision tools for a better management of resources and risks. Extreme value theory (EVT) is a branch of statistics that provides a suitable framework for the modeling of such events. Besides, the distribution of precipitation is non-stationary in space owing to the spatial heterogeneity of the Mediterranean area. To take into account spatial non-stationarity, the parameters of the distribution can be considered as functions of covariates. These functions may be implemented with flexible non-parametric non-linear models such as Artificial Neural Networks (ANN). The main issue in this analysis is the selection of the level of complexity of the ANN (i.e. the number of hidden neurons) and the optimal choice of covariates used as inputs to the ANN. Three sites are considered to form a north-south aridity gradient. The results presented mainly concern the southernmost site in central Tunisia.

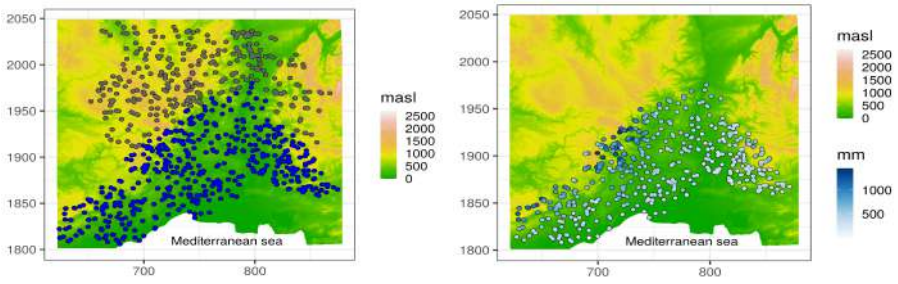
Keywords. Intense precipitation events, spatial non-stationarity, extreme value theory, artificial neural networks.

1 Introduction

The Mediterranean region is known for its critical vulnerability to climate change. Among hydrometeorological processes, we are interested in the study of extreme precipitation events. The irregularity of precipitation events and their high spatiotemporal variability lead to problems of water resources scarcity. Thus, rainfall scenarios with high spatiotemporal resolution are necessary to implement decision support tools for better management of resources and risks. The main goal of this work is to put forward a spatial interpolation analysis to characterize intense precipitation events in a Mediterranean context.

2 Presentation of study sites

Three Mediterranean sites that form a north-south aridity gradient are included in our study: part of the French Mediterranean, the Cap Bon in North-East Tunisia and the Merguellil catchment in central Tunisia. Daily rainfall totals are available for each regions. To reduce the effect of spatial heterogeneity, we identify stations in the French region sharing a Mediterranean climate by K-means clustering. Each station is characterized by a vector of standardized high quantiles of annual maxima together with altitude and latitude. The stations retained as Mediterranean are indicated in blue in Figure 1a.

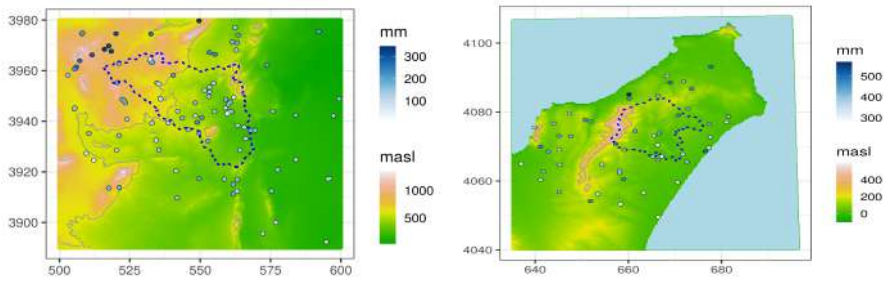


(a) Stations with a Mediterranean climate in blue.

(b) Average cumulative precipitation over the rainy season (October-April).

Figure 1: French Mediterranean region.

After preliminary analyzes and building on previous work, e.g. Tramblay and Hertig (2018), we define the rainy season for this region from October to April. For the two Tunisian regions, all the stations are kept since the region is smaller and does not include more than one type of climate. The rainy season for both Tunisian regions is the period from October to March (Mekki I. 2003; Lacombe G. 2007). In this work, temporal non-stationarity modelling is bypassed by sampling only over months that belong to the rainy season. In this case, we will assume that temporal non-stationarity does not need to be managed further.



(a) Merguellil (b) Lebna, Cap Bon

Figure 2: Average cumulative precipitation during the rainy season (October-March) in the Tunisian regions.

3 Methodology

3.1 Extreme value theory

Extreme Value Theory (EVT) is the branch of statistics that provides suitable tools for extreme value analysis (Coles, 2001). We are interested to estimate the behavior of the upper tail of the distribution that concerns large events. Suppose that $X_1 \dots X_n$ is a sequence of independent random variables from F a distribution function. $M_n = \max\{X_1 \dots X_n\}$ is the maximal value of the random variables. The fundamental theorem of Fisher and Tippett gives an asymptotic result to characterize the maximum of a sample. Suppose that there exist $c_n > 0$ and $d_n \in \mathbb{R}$ two sequences of real numbers, such that:

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leq x\right) = G(x) \tag{1}$$

where G is a non-degenerate distribution function that can be expressed as the generalized extreme value (GEV) distribution:

$$G(z) = \exp -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi} \tag{2}$$

defined for the values of z for which $1+\xi(z-\mu)/\sigma > 0$ and (μ, σ, ξ) are the location, scale and shape parameters, respectively. Each distribution corresponds to a different behavior of the maximum controlled by the shape parameter ξ.

The block maxima approach consists in sampling the observed maximum over a block, often chosen to correspond to a period of one year. To summarize extreme behavior from the distribution function, we can consider the upper quantiles of a high order. To estimate these quantiles for the GEV, we invert equation (2) such that $G(z_p) = 1-p$ to get a return level z_p which should be exceeded by the annual maximum with a probability p.

3.2 Spatial non-stationarity for extremes

The spatial and temporal variability in the region is explained by the presence of mountainous and hilly landforms and by the maritime and desert influences depending on the wind fields. This is why the previous assumption of stationarity of the annual maxima M_n is not true in practise. To take into account spatial non-stationarity, we can consider that the parameters of the model vary as a function of geographical covariates (Blanchet and Lehning 2010). We propose to use an artificial neural network (ANN) with a feed-forward architecture to implement the function that estimates the GEV parameters. The ANN has the ability to represent flexible relationships between input and output variables. It can model spatial non-stationarity by considering that the parameters of the distribution are the outputs of the network that vary in space as a function of covariates.

Let $x = \{x_1, \dots x_n\}$ be the input variables (or covariates). For each neuron j with an activation function h, transforms m linear combination of the inputs and gives as output:

$$z_j = h(\sum_{i=1}^{n} w_{ji} x_i + w_{j0}), \tag{3}$$

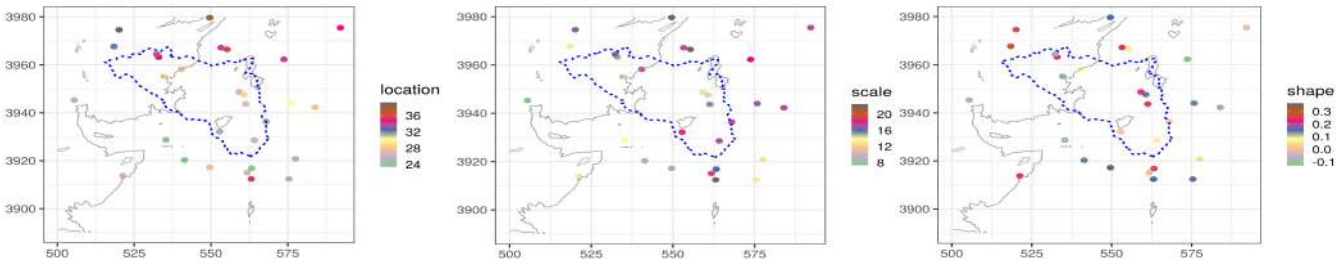
Where $j = 1, \dots, m$. w_{ji} is the weight of the hidden layer connecting the neuron j with the i^{th} component of the input vector, and w_{j0} is the bias of the neuron j (i.e. a constant). The function $h(.)$ is a nonlinear activation function usually taken as the hyperbolic tangent. The z_j are linearly combined to give the activation of the output units, with activation function σ and with $k = 1 \dots K$, K is the total number of outputs :

$$y_k = \sigma(\sum_{j=1}^{m} w_{kj} z_j + w_{k0}) \tag{4}$$

4 Preliminary results

As a first step, the GEV distribution parameters, see (2), for each region are estimated based on maxima per rainy season at each station separately. $M_n(s)$ is a rainy season maxima for station s on year n which is assumed to follow a GEV distribution with the parameters (μ_s, σ_s, ξ_s) to be estimated with the L-moments method. In Figure 3, the resulting GEV parameter estimates are presented for the Merguellil catchment.

As a second step, the 20 year return levels computed from the local GEV parameters estimated at the first step are interpolated spatially. To this end, we use an ANN as described in 3.2. The main issue with ANN is the choice of the number of hidden units. With too few hidden units, the ANN is biased and the fitting is not achieved because it does not have enough flexibility. In contrast, with a large number of units, the fitting becomes very sensitive to the variance, i.e. it reaches the overfitting phenomenon. The solution is to check a compromise between these two aspects: underfitting and overfitting.

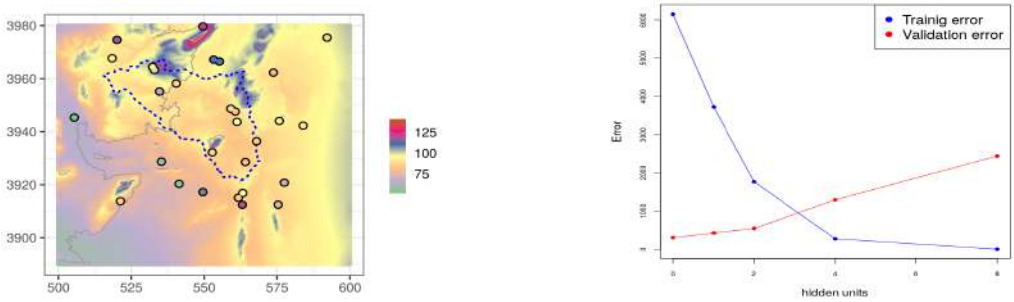


(a) Location (b) Scale (c) Shape

Figure 3: Estimation of GEV parameters by station.

This is achieved by selecting the number of hidden units corresponding to the lowest validation error as computed by the leave-one-out scheme, also known as jackknife.

With the covariate set fixed to the x-y-z coordinates, we applied the leave-one-out scheme to select the optimal number of hidden units for this particular covariate set. We obtained the graph of training and validation error (Figure 4b). In this case, the selected number of hidden units is $nh = 0$. The corresponding spatial interpolation of the 20 year return levels is presented in Figure 4a.

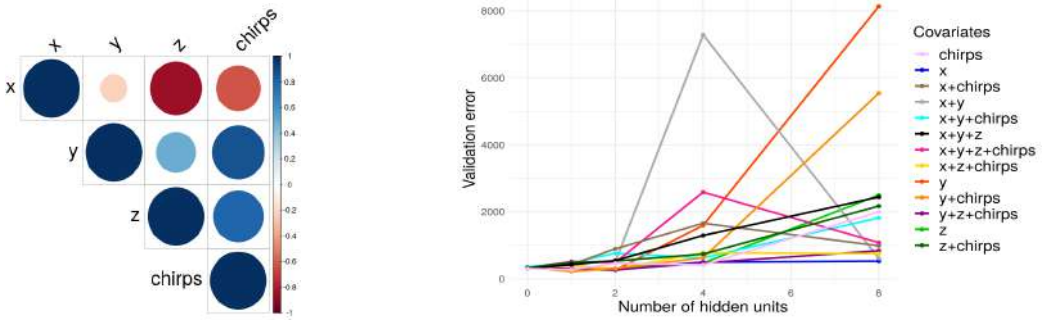


(a) Interpolated 20 year return levels. (b) Training and validation errors.

Figure 4: Merguellil catchment : spatial interpolation of the 20 year return levels with a fixed set of geographic covariates (x, y, z). The training and validation errors are computed with the leave-one-out scheme for increasing number of hidden units.

To select an optimal covariate set, the above procedure is repeated for several candidate covariate sets. We propose to include the average precipitation totals during the rainy season as an indication of climatic variability. This information is computed from the CHIRPS reanalysis daily precipitation dataset. The climatic covariate derived from CHIRPS is correlated with the x, y and z coordinates according to the Kendall correlation coefficients in Figure 5a. The leave-one-out scheme was applied to different combinations of these four covariates and presented in Figure 5b.

The common choice of the optimal number of hidden units and the appropriate covariates corresponds to the value of the minimum validation error. In our case, an ANN with a single hidden unit using as covariates the y-coordinate and CHIRPS climatic covariate is the optimal choice.



(a) Kendall correlation for $(x, y, z, chirps)$.

(b) Validation errors for each candidate covariate sets.

Figure 5: Selection of the covariate set with the leave-one-out scheme.

5 On-going work

In order to translate properly the spatial non-stationarity for extreme distributions, we propose to consider that the parameters of the GEV vary as a function of covariates directly (with the so-called surface responses) rather than going through the punctual estimation. Once we get to fit an ANN with the appropriate covariates, we aim to make a comparison with a generalized linear model.

Bibliography

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics, Springer-Verlag, London.

Tramblay, Y. and Hertig, E., (2018), Modelling extreme dry spells in the Mediterranean region in connection with atmospheric circulation. Atmos Res 202:40–48.

Mekki, I. (2003), Analyse et modélisation des flux hydriques à l'échelle d'un bassin versant cultivé alimentant un lac collinaire du domaine semi-aride méditerranéen (Oued Kamech, Cap Bon, Tunisie). Thèse de doctorat, Université Montpellier II, pp : 170.

Lacombe, G. (2007), Evolution et usages de la ressource en eau dans un bassin versant aménagé semi-aride : le cas du Merguellil en Tunisie centrale. Thèse de doctorat, Université Montpellier II, IRD, 304 p. multigr. Th. : Eaux Continentales et Société.

Blanchet, J., and Lehning, M. (2010). Mapping snow depth return levels: Smooth spatial modeling versus station interpolation. Hydrology and Earth System Sciences,14(12), 2527–2544.

Impacts calculation and visualization in spatial flows modeling, application to remittances

Thibault Laurent [1] & Paula Margaretic [2] & Christine Thomas-Agnan [3]

[1] *Toulouse School of Economics (CNRS), 1, Esplanade de l'Université, 31080 Toulouse Cedex 06, FRANCE, Thibault.Laurent@tse-fr.eu*

[2] *Universidad de San Andrés, Vito Dumas 284, Victoria, B1644BID Buenos Aires, ARGENTINA, pmargaretic@udesa.edu.ar*

[1] *Toulouse School of Economics (Université Toulouse 1 Capitole), 1, Esplanade de l'Université, 31080 Toulouse Cedex 06, FRANCE, Christine.Thomas@tse-fr.eu*

Résumé. Un flux spatial correspond à un transfert entre deux unités géographiques (une origine et une destination) de différentes quantités socio-démographiques (déplacements de population de type domicile-travail, échanges monétaires, envois de fonds, etc.). Pour modéliser de telles données, il faut tenir compte des caractéristiques observées à la fois à l'origine et à la destination (O-D), ainsi que des covariables propres aux flux eux-mêmes comme les distances géographiques entre origine et destination. Le modèle de base est le modèle gravitaire qui peut être biaisé si les résidus sont spatialement autocorrélés. Dans ce travail, on considère le modèle d'interaction spatial (SIM) (LeSage et Pace, 2008) et on présente une nouvelle procédure de décomposition des impacts marginaux (décomposition en origine/destination/intra/réseau/total dans la lignée de LeSage et Thomas-Agnan, 2014, 2015). On s'intéresse également à la visualisation de ces impacts à l'aide de graphiques statistiques. Le travail sera illustré par une application empirique liée aux envois de fonds entre pays.

Mots-clés. Statistique spatiale, économétrie spatiale

Abstract. A spatial flow corresponds to a transfer between two geographical units (an origin and a destination) of various socio-demographic quantities: population, monetary exchanges, remittances, etc. For modelling such data, one needs to take into account the characteristics observed at both origin and destination (OD) and also the covariates related to the flows themselves like the geographical distances between O and D. The basic model is the gravity model whose parameters can be biased if the residuals are spatially autocorrelated. LeSage and Pace (2008) define the Spatial Interaction Models (SIM). We present a new procedure (in line with LeSage and Thomas-Agnan, 2014, 2015) to decompose the marginal impacts into origin, destination, intra, network, and total effect and several statistical graphics for visualising these impacts. The work will be illustrated by an empirical application related to the remittances between countries.

Keywords. Spatial statistic, spatial econometrics.

1 Introduction

Flow data are frequent in regional science and represent movements of intangible entities (such as information or knowledge) or tangible ones, such as people, goods or remittances between two spatial locations. For instance, the top part of Figure 1 represents a Sankey diagram for the remittances between geo-economic zones in 2017 and the bottom part of Figure 1 represents a heatmap of the origin-destination (OD hereafter) matrix with remittances flows, by country (Source : World Bank).

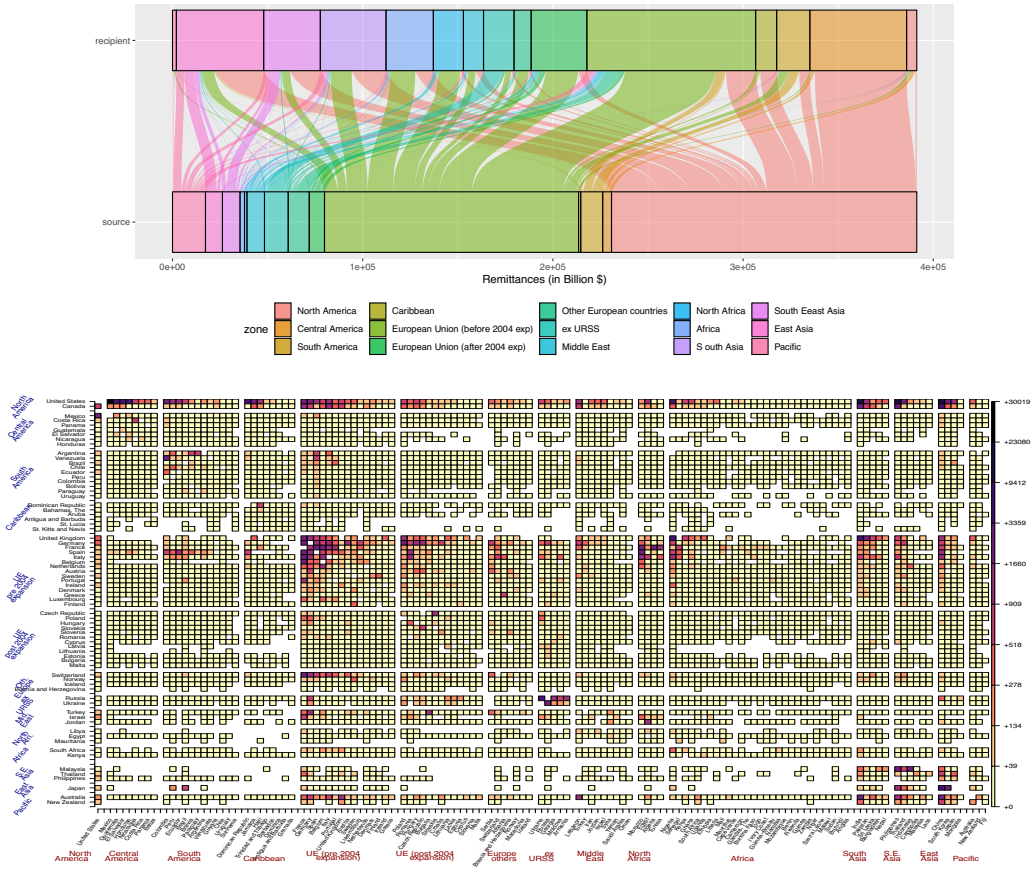


Figure 1: On the top, Sankey diagram of the remittances flows, by geo-economic zones, on the bottom, OD heatmap of the remittance flows, by country (World Bank, 2017)

Gravity models have been extensively used to model OD flow data, with applications in migration, international trade, demography, tourism and transportation, among other fields. In this work, we use the spatial interaction models (SIM hereafter) popularized by LeSage and Pace (2008). LeSage and Thomas-Agnan (2014, 2015) demonstrate that, in

the case of the SIM, the estimated parameters cannot be directly interpreted as marginal effects or elasticities, and that the bidimensionality of the SIM, where both origin and destination-level effects are often estimated for a variable, introduces further complications. Besides, they show how to calculate the marginal effects of explanatory variables in the framework of the spatial interaction autoregressive model.

They distinguish between the origin effects, destination effects, intraregional effects, and network/spillover effects, arising from changes in the characteristics/explanatory variables. Since they propose scalar summary measures for the four types of effects, their development allows the interpretation of estimates in a manner similar to that used in typical regression models.

Two common assumptions in the spatial interaction model applications computing impact measures are that there is the same set of explanatory variables for the origins and the destinations and the same list of locations, for both the origins and destinations. Therefore, in this paper, we extend LeSage and Thomas-Agnan (2015) by relaxing the previous two assumptions. Precisely, we allow for possibly different subsets of origin and destination characteristics and for the possibility of a different list of locations for the origins and destinations.

To illustrate our application, we rely on a sample of bilateral remittances (coming from the World Bank). Our dataset comprises 67 source countries and 129 recipient countries all over the world, for which workers' remittances are reported in 2017.

2 General specification of a spatial interaction model

Let Y be the flow matrix, where the n_o columns represent the origins 1 to n_o, the n_d rows correspond to destinations 1 to n_d and $N = n_o \times n_d$ is the total number of OD flows:

$$Y = \begin{pmatrix} o_1 \to d_1 & o_2 \to d_1 & ... & o_{n_o} \to d_1 \\ o_1 \to d_2 & o_2 \to d_2 & ... & ... \\ & & & o_{n_o} \to d_{n_d-1} \\ & & & o_{n_o} \to d_{n_d} \end{pmatrix} \tag{1}$$

We denote by $Y_{i:j}$ an OD flow from origin i to destination j. Two possible vectorizations of the flow matrix Y are possible, depending on whether we stack the columns (destination centric) or the rows (origin centric) of the flow matrix. In this paper, we choose a destination centric ordering and denote by y, the flow vector, of length $N \times 1$. Hence, the first n_d elements of y represent flows from origin 1 to all n_d destinations. All formulas below can to be adapted to the origin-centric scheme.

To write the model, it is convenient to use the Kronecker product $\bigotimes$ to express some vectors and matrices, as it has the computational advantage to avoid storing multiple copies of the same numerical value. For example, the characteristic of the origin i will appear as an explanatory variable in all flows originating from i.

Given a neighbourhood structure for the set of origins and destinations, several types of neighborhood structures can be defined for the flows (see LeSage and Pace, 2008), namely, structures capturing origin-based dependence, destination-based dependence and/or origin-destination based dependence. To implement them, we define: OW of dimension $n_o \times n_o$ for characterizing the proximity in the set of origins and DW of dimension $n_d \times n_d$ for characterizing the proximity in the set of destinations. We can then obtain the three types of neighborhood structures as follows: $W_o = OW \bigotimes I_{n_d}$ is the origin based spatial neighborhood matrix, $W_d = I_{n_o} \bigotimes DW$ is the destination based spatial neighborhood matrix, $W_w = OW \bigotimes DW$ is the origin-to-destination based spatial neighborhood matrix. Note that the three weight matrices W_o, W_d and W_w are of dimension $N \times N$.

We have two sets of characteristics for the spatial units: One set characterizing the locations in the set of origins, and a second set characterizing the locations in the set of destinations. For example, in the analysis of home to work commuting data, the factor population size of a given area is a determinant of the flows originating from this location, whereas the number of jobs in a given area is a determinant of the flows into this location.

Let OX be the matrix of origin characteristics and DX that of destination characteristics. To allow for the Durbin aspect, some of the origin (respectively, destination) characteristics may appear in their lagged form in the model and we denote them by OLX (respectively, DLX). We now construct the following four matrices: $X_o = OX \bigotimes \iota_{n_d}$, characteristics of the spatial units which act as origins, $X_d = \iota_{n_o} \bigotimes DX$, characteristics of the spatial units which act as destinations, $XL_o = OLX \bigotimes \iota_{n_d}$, lagged characteristics of the spatial units acting as origins, $XL_d = \iota_{n_o} \bigotimes DLX$, lagged characteristics of the spatial units acting as destinations.

Let G be the matrix of variables characterizing both origin and destination (distance, for example). Finally, the model in its reduced form can be written as

$$(I_{N\times N} - \rho_o W_o - \rho_d W_d + \rho_w W_w)Y = X_o\beta_o + X_d\beta_d + XL_o\delta_o + XL_d\delta_d + G\gamma + \epsilon, \quad (2)$$

where β_o, β_d, δ_o, δ_d and γ are vectors of parameters whose dimension correspond to the number of variables in OX, DX, OLX, DLX and G, respectively.

Let $A(W) = (I_{N\times N} - \rho_o W_o - \rho_d W_d + \rho_w W_w)^{-1}$ be the $N \times N$ filter matrix. It could be another filter for one of the nine flow submodels described in LeSage and Pace (2008). To estimate the parameters of the model (2), we use the Bayesian and MCMC approach (LeSage and Pace, 2009).

3 Impacts in the Spatial Interaction Durbin model

As Pace and LeSage (2008) point out, the classical interpretation of coefficients as marginal effects in least square models cannot be directly extended to all spatial models. In an ordinary linear model, the coefficient β_r of a particular variable x_r corresponds to the partial derivative of the i-th component $\mathbb{E}(y_i)$ of the expected dependent variable with

respect to x_{ir} for any i (and hence constant across i). It is usually interpreted as the increment of $\mathbb{E}(Y)$ when the r-th explanatory variable increases by one unit (each component), all other variables being held fixed. Note also that the partial derivative of the i-th component $\mathbb{E}(y_i)$ of the expected dependent variable with respect to x_{jr} for $j \neq i$ is equal to zero.

In the LAG model, it is easy to see that the cross partial derivative of the i-th component $\mathbb{E}(y_i)$ with respect to x_{jr} for $j \neq i$ is not anymore equal to zero, implying that the change of an explanatory variable for the i-th location will affect not only y_i but all the y_j. Moreover, the effect on $\mathbb{E}(y_i)$ of increasing by one unit the j-th component of an explanatory variable x_{jr} is no longer constant across locations i (see example in figure 2).

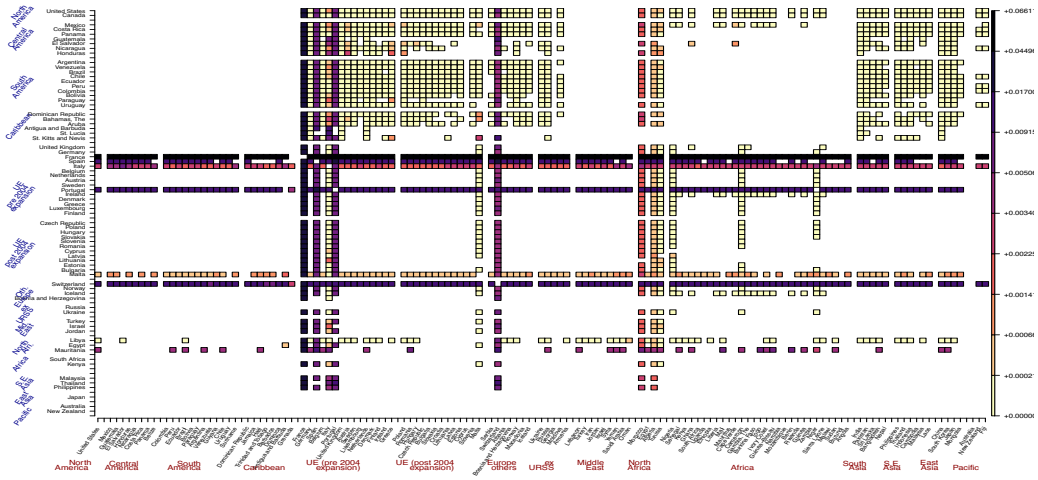


Figure 2: Impacted flows due to a change of x in France

In a spatial interaction model, LeSage and Thomas-Agnan (2015) show that changing a characteristic may have an intra-regional effect, an origin effect, a destination effect, an indirect effect and an overall effect. Let $R^r_{i:j,t} := \frac{\partial \mathbb{E}(y_{i:j})}{\partial x_{tr}}$ be the marginal impact of variable r measured at location t on the flow from i to j. Then these effects are expressed as follows: $R^r_{i:j,t} = \sum_{k=1}^{n_o} A_{i:j,k:t}\beta_o + \sum_{l=1}^{n_d} A_{i:j,t:l}\beta_d$

1. Mean origin effect of $X_{.r}$: $OE^r = \frac{1}{n_o} \sum_{k=1}^{n_o} \sum_{j=1,j\neq k}^{n_d} R^r_{k:j,k}$
2. Mean destination effect of $X_{.r}$: $DE^r = \frac{1}{n_d} \sum_{k=1}^{n_d} \sum_{i\neq k,i=1}^{n_o} R^r_{i:k,k}$
3. Network effect of $X_{.r}$: For the network effect, we average over all possible locations t; therefore, the normalizing factor will be $m = n_o$ if X is an origin characteristic and $m = n_d$ if X is a destination characteristic: $NE^r = \frac{1}{m} \sum_{k=1}^{m} \sum_{i\neq k,i=1}^{n_o} \sum_{j\neq k,j=1}^{n_d} R^r_{i:j,k}$

Finally, if X acts at the same time as an origin characteristic and as a destination characteristic, we will need to add the two effects thus obtaining, $NE^r = \frac{1}{n_d}\sum_{k=1}^{n}\sum_{i\neq k}\sum_{j\neq k} R^r_{i:j,k} + \frac{1}{n_o}\sum_{k=1}^{n}\sum_{i\neq k}\sum_{j\neq k} R^r_{i:j,k}$

4. Total effect of $X_{.r}$: $TE^r = DE^r + OE^r + NE^r$

Different representations of the impacts can be done. For instance, Figure 3 represents the OD, DE and NE group by country due to a change of the explanatory variable GDP per capita.

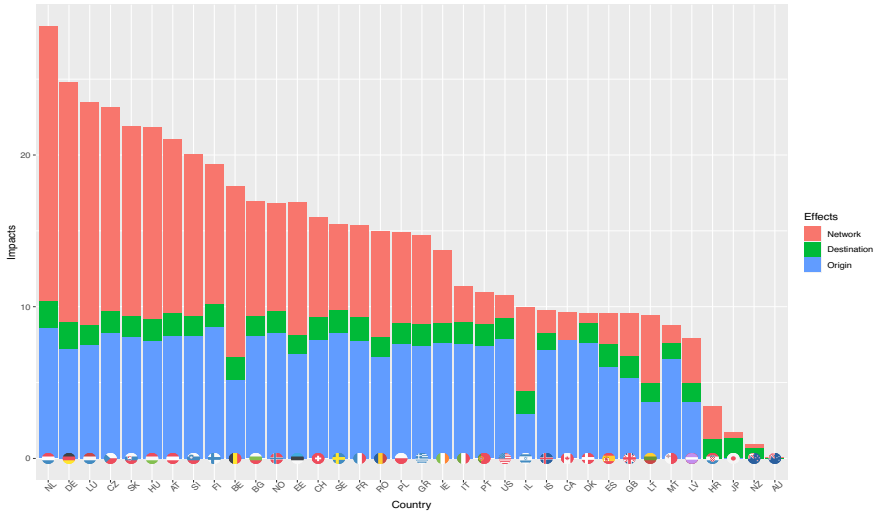


Figure 3: Summarizing the impacts by country

References

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

LeSage, J. P. and Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48(5):941–967.

LeSage, J. P. and Thomas-Agnan, C. (2014). Spatial econometric OD-Flow models. In Fischer, M. M. and Nijkamp, P., editors, *Handbook of Regional Science*, pages 1653–1673. Springer.

LeSage, J. P. and Thomas-Agnan, C. (2015). Interpreting spatial econometric origin-destination flow models. *Journal of Regional Science*, 55(2):188–208.

Générateur d'Euler conditionnel pour les séries chronologiques

Carl REMLINGER [1,2] & Joseph MIKAEL [2,3] & Romuald ELIE [1]

[1] *Université Gustave Eiffel*
[2] *EDF Lab*
[3] *FiME (Laboratoire de Finance des Marchés de l'Energie)*

Résumé. Nous proposons un générateur de séries temporelles reposant sur une discrétisation d'Euler. Plus précisément, nous développons un générateur Euler conditionnel (CEGEN) qui minimise une distance entre les distributions conditionnelles induites par les processus. Dans le cas de processus d'Itô, nous montrons qu'en utilisant la métrique de Bures, atteindre une faible erreur fournit une estimation précise des termes de drift et de volatilité des processus sous-jacents. Des tests sur des processus usuels mettent en évidence comment la discrétisation d'Euler et l'utilisation de la distance de Wasserstein permettent à notre modèle de surpasser l'état de l'art des générateurs de séries temporelles Yoon et al. (2019). Nous observons qu'en grande dimension CEGEN retrouve les bonnes structures de covariance. Enfin, nous illustrons comment notre modèle peut être combiné à un simulateur Monte Carlo en utilisant le transfer learning dans un contexte où il y a peu de données.

Mots-clés. série temporelle, modèle génératif, distance de Wasserstein, métrique de Bures

Abstract. We propose a generator for time series based on an Euler discretization. More precisely, we develop a conditional Euler generator (CEGEN) which minimizes a distance between the conditional distributions induced by the processes. In the case of Itô processes, we show that using the Bures metric, achieving a low error provides an accurate estimate of the drift and volatility terms of the underlying processes. Tests on usual processes highlight how Euler's discretization and the use of Wasserstein distance allow CEGEN to surpass the state of the art of time series generators Yoon et al. (2019). We observe that in large dimension CEGEN finds the good covariance structures. Finally, we illustrate how our model can be combined with a Monte Carlo simulator using transfer learning in a context where there is little data.

Keywords. Time series, Generative model, Wasserstein distance, Bures Metric

1 Introduction

Les simulations Monte Carlo de séries temporelles sont largement utilisées dans diverses applications industrielles comme la prise de décision, le contrôle stochastique, ou

encore la prédiction. Elles sont abondamment employées dans le secteur financier pour les stress tests du marché Sorge (2004), le contrôle des risques et la couverture, Buehler et al. (2019) et Fécamp et al. (2020), ou pour la mesure d'indicateurs de risque usuels comme la Value-at-Risks (Jorion (2000) par exemple).

La conception d'un modèle réaliste et interprétable reste une tâche fastidieuse et principalement manuelle, qui nécessite des hypothèses de modélisation sur la dépendance temporelle des variables. Il est nécessaire de sélectionner un type de modèle, tels que les modèles ARMA, GARCH, Black-Scholes ou Heston qui sont des références très courantes dans l'industrie. Il n'est donc pas simple de mettre à jour ces modèles lorsque des données d'un nouveau type sont observées, par exemple l'apparition de prix négatifs sur les marchés de l'électricité, une crise économique ou sanitaire modifiant la structure du marché ou de nouvelles conditions météorologiques. Cela amène naturellement au développement de générateurs plus souples et fiables pour les séries chronologiques.

Les méthodes génératives telles que Variational Auto Encoders (VAE) de Kingma et Welling (2013) et Generative Adversarial Networks (GAN) de Goodfellow et al. (2014) affichent des résultats prometteurs pour les données financières, voir Xu et al. (2018). Le développement de méthodes génératives similaires appliquées aux séries temporelles est particulièrement attrayant car il apprendrait directement un modèle de diffusion à partir des données, sans hypothèses de modélisation sous-jacentes. Cependant, en raison de la structure temporelle complexe et éventuellement non stationnaire de la série chronologique initiale, de telles méthodes génératives sont difficiles à appliquer en tant que telles. La figure 1 représente deux processus générés et illustre l'une des difficultés : le

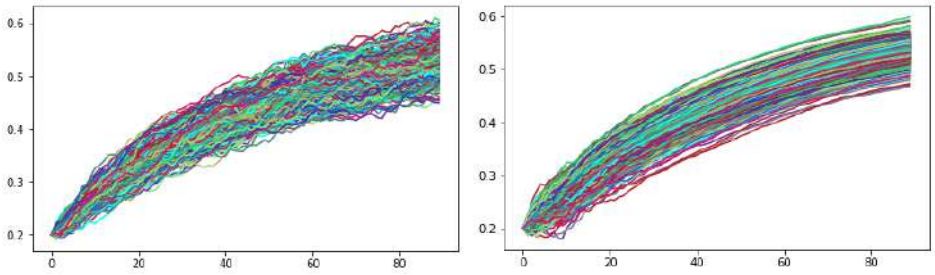


FIGURE 1 – À **droite**, un ensemble de mauvaises trajectoires générées où à chaque date les marginales correspondent avec le processus original affiché à **gauche**.

générateur est capable d'apprendre correctement les distributions marginales à chaque date, mais n'arrive pas à capturer la dynamique du processus. Une modélisation efficace de séries chronologiques apparaît donc comme un équilibre entre retrouver les marginales et capturer la structure temporelle.

2 Contributions

Pour répondre à ces difficultés, nous construisons des générateurs reposant sur un schéma d'Euler en nous concentrant sur la structure générale des processus d'Itô. Cette représentation nous permet de fournir une formulation mathématique rigoureuse du problème et d'envisager des contraintes pertinentes sur le générateur. L'utilisation de ce schéma pourrait également être généralisé en ajoutant des sauts ou des structures d'autocorrélation. Nous introduisons un nouveau générateur (CEGEN) reposant sur les densités conditionnelles et la distance de Wasserstein-2, capable de rivaliser avec l'état de l'art Time Series GAN (TSGAN) proposé par Yoon et al. (2019). Nous fournissons une preuve théorique qu'une bonne précision sur la fonction de perte conditionnelle introduite implique une estimation correcte des coefficients du processus d'Itô. Des tests numériques comparant l'approche GAN et celle conditionnelle sont réalisés jusqu'à la dimension 20, et montrent de bons résultats à la fois sur les métriques de structure temporelle et sur les métriques de distribution marginale. Enfin, nous démontrons la robustesse des performances de notre générateur avec une situation réaliste. Un praticien n'a pas assez de données pour entraîner correctement un générateur mais a un modèle cohérent A qui doit être amélioré avec des données historiques. Nous illustrons comment notre générateur peut utiliser le transfer learning en commençant son apprentissage sur A et en affinant ses paramètres avec une faible quantité de données historiques.

3 Résultats

Afin de permettre une formulation mathématique solide qui ouvre un large choix possible d'applications, nous représentons notre série temporelle comme un processus d'Itô $X = (X_{t_i})_{i=1..N} \in \mathbb{R}^{d\times N}$ observé sur l'espace $t_0 < t_1 < ... < t_N$, avec $\Delta t := t_{i+1} - t_i$.

$$dX_t \quad = \quad b_X(t, X_t)dt + \Sigma_X(t, X_t)dW_t,$$

où $b : \mathbb{R}^{d+1} \to \mathbb{R}^d$, $\Sigma_X : \mathbb{R}^{d+1} \to \mathcal{M}_{d\times d}$ et W est un Brownien de dimension d. La discrétisation d'Euler en fonction du temps (sur les $(t_i)_i$) du processus ci-dessus est donnée par :

$$X_{t_i+\Delta t} \quad = \quad X_{t_i} + b_X(t_i, X_{t_i})\Delta t + \Sigma_X(t_i, X_{t_i})\Delta W^1_{t_i}$$

où $\Delta W^i_t \sim \mathcal{N}(0, \Delta t I_d)$.

Nous cherchons un générateur g_θ qui simule une série temporelle aléatoire Y à la date t_i étant la plus proche de X en terme de distribution et de dynamique. Nous proposons de concevoir un générateur, reposant sur sur des densités conditionnelles, qui modélise les coefficients de drift b_Y et de volatilité Σ_Y qui apparaissent dans l'équation suivante :

$$Y_{t_i+\Delta t} \quad = \quad Y_{t_i} + g^b_\theta(t_i, Y_{t_i})\Delta t + g^\Sigma_\theta(t_i, Y_{t_i})\Delta W^2_{t_i}$$

Cette modélisation permet de construire un algorithme pratique car nous pouvons comparer directement les drifts et les volatilités dans les cas où b_X et Σ_X sont connus. Nous précisons que le générateur g_θ n'a jamais connaissance des vrais coefficients b_X and Σ_X durant l'entraînement.

La difficulté qui se pose lorsque nous essayons d'appliquer des générateurs à des séries temporelles vient de la nécessité de trouver le juste équilibre entre l'estimation marginale et la structure temporelle. Avec le conditionnement nous parvenons à obtenir un modèle qui allie ces deux propriétés avec peu d'hypothèses :

Proposition 3.1 *Soient $\varepsilon > 0$ et $z \in \mathbb{R}^d$ tel que $z^j \in [a_k^j, a_{k+1}^j]$ avec $a_k^j < a_{k+1}^j$. Soient b_X, b_Y deux fonctions K-Lipshitz par rapport à la seconde variable, et Σ_X,Σ_Y deux fonctions strictement positives et K-Lipshitz par rapport à la seconde variable. Supposons que pour tout $t_i \in \{t_0, \dots, t_N = T\}$, $\Delta t = t_{i+1} - t_i$,*

$$\mathcal{W}_2\left(\mathcal{L}(X_{t_{i+1}}|(X_{t_i} \in [a_k, a_{k+1}])) \, , \, \mathcal{L}(Y_{t_{i+1}}|(Y_{t_i} \in [a_k, a_{k+1}])\right) \leq \varepsilon$$

Alors,

$$\|b_X^j(t_i, z)\Delta t - b_Y^j(t_i, z)\Delta t\|_2^2 \leq \varepsilon + (2K+1)\|a_{k+1}^j - a_k^j\|_2^2$$

De plus,

- *si $d = 1$, $\|\Sigma_X(t_i, z) - \Sigma_Y(t_i, z)\|_2^2 \leq \varepsilon + 2K\|a_{k+1} - a_k\|_2$.*
- *si $d > 1$ et $Tr(\Sigma_X(t_i, z)\Sigma_X(t_i, z)^T) = Tr(\Sigma_Y(t_i, z)\Sigma_Y(t_i, z)^T) = \alpha$, alors*

$$\|\Sigma_X^j(t_i, z) - \Sigma_Y^j(t_i, z)\|_2^2 \;\leq\; \frac{\sqrt{2}\varepsilon}{\alpha\sqrt{\Delta t}} + 2K\|a_{k+1}^j - a_k^j\|_2^2.$$

L'approche supervisée de la méthode conditionnelle CEGEN, qui cherche à minimiser la distance de Wasserstein entre les lois conditionnelles, surpasse le modèle semi-supervisé TSGAN autant sur le plan des marginales que sur la structure temporelle, comme le montre la figure 2. La variance entre les incréments ou la distance de Fréchet Heusel et al. (2017) favorise le choix du générateur CEGEN. Des échantillons de trajectoires sont affichés figure 3.

La situation réaliste évoquée en introduction met aussi en avant le modèle CEGEN. Souvent, nous n'avons pas suffisamment de données pour entraîner un générateur. Lorsqu'un modèle raisonnable est déjà disponible, nous pouvons utiliser des techniques d'augmentation de données, via des méthodes de transfer learning, afin de commencer l'apprentissage du générateur. Ensuite pour affiner le modèle, nous l'entraînons avec des données historiques. Nous comparons les quantiles des échantillons générés et observons qu'en alliant apprentissage par transfert et le générateur CEGEN, ce dernier surpasse le simulateur de Monte Carlo.

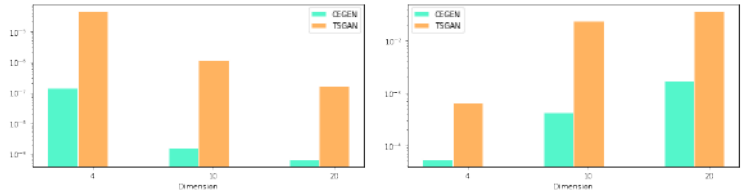


FIGURE 2 – **Gauche :** Moyenne sur le temps de la distance de Fréchet entre les distributions marginales. **Droite :** Moyenne sur le temps de la norme ℓ_2 entre les variances de dX_t et celles de dY_t.

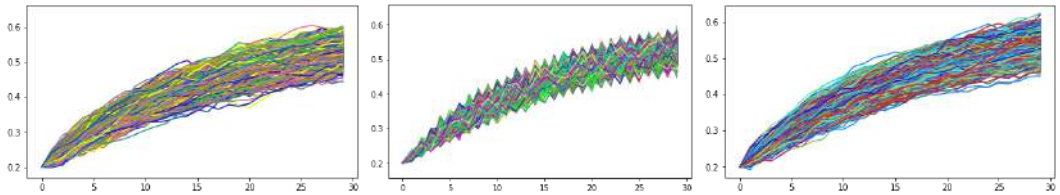


FIGURE 3 – **Droite :** Simulations Monte Carlo. **Milieu :** Trajectoires générées par TS-GAN. **Droite :** Trajectoires générées par CEGEN.

4 Conclusion

Un générateur pour séries temporelles reposant sur le schéma d'Euler est proposé. En utilisant les distributions conditionnelles le générateur montre de bons résultats en terme de représentation des séries et résiste à la grande dimension. L'apprentissage par transfert pour cette méthode met en évidence la souplesse et la rapidité du modèle à s'adapter à un changement de régime. La formulation mathématique que nous proposons permet de montrer qu'une erreur suffisamment faible donne une représentation fidèle des séries chronologiques. Notre algorithme peut donc générer fidèlement des processus d'Itô et permet un contrôle sur les termes de drift et de volatilité. La classe de ces processus est suffisamment large pour modéliser la plupart des séries d'intérêt et pourrait être étendue dans de prochains travaux à la classe des processus de Lévy.

Bibliographie

Buehler, H. et al. (2019). "Deep hedging". In : *Quantitative Finance* 19.8, p. 1271-1291.

Fécamp, S. et al. (2020). "Deep learning for discrete-time hedging in incomplete markets". In :

Goodfellow, I. et al. (2014). "Generative adversarial nets". In : *Advances in neural information processing systems*, p. 2672-2680.

Heusel, M. et al. (2017). “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In : *arXiv preprint arXiv :1706.08500*.

Jorion, P. (2000). *Value at risk*. McGraw-Hill Professional Publishing.

Kingma, D. P. et M. Welling (2013). “Auto-encoding variational bayes”. In : *arXiv preprint arXiv :1312.6114*.

Sorge, M. (2004). “Stress-testing financial systems : an overview of current methodologies”. In :

Xu, J. et al. (2018). “DP-GAN : diversity-promoting generative adversarial network for generating informative and diversified text”. In : *arXiv preprint arXiv :1802.01345*.

Yoon, J. et al. (2019). “Time-series generative adversarial networks”. In :

Détection d'individus atypiques en régression SIR (sliced inverse regression)

Hadrien Lorenzo [1,2] & Jérôme Saracco [1,2,3]

[1] *ASTRAL, INRIA BSO, 200 Avenue de la Vieille Tour, 33405 Talence, France*
[2] *Institut de mathématiques de Bordeaux, UMR 5251 CNRS, 33400 Talence, France*
[3] *ENSC - Bordeaux INP, 33400 Talence, France*

Résumé. La régression inverse par tranches (*sliced inverse regression*, SIR) considère un modèle semi-paramétrique de régression entre une variable dépendante y et une variable explicative p-dimensionnelle x via un indice $x'\beta$ et une fonction de lien f. La direction du paramètre p-dimensionnel β, appelée direction EDR (pour *effective dimension reduction*), est estimée via la méthode SIR. Le paramètre fonctionnel f peut être estimé par un estimateur à noyau de la régression en utilisant l'indice estimé. Cependant, si des observations atypiques (*outliers*) sont présentes dans les données, cette méthodologie en deux étapes ne va plus fonctionner correctement. L'objet de cette communication est de présenter trois méthodes computationnelles permettant de détecter les individus atypiques dans ce modèle semi-paramétrique de régression. Le comportement numérique de ces méthodes (implémentées dans R) est illustré avec des simulations. Un exemple sur données réelles est également présenté.

Mots-clés. Modèle semi-paramétrique de régression, régression inverse par tranches (sliced inverse regression, SIR), estimateur non-paramétrique à noyau d'une courbe de régression, détection d'individus atypiques/aberrants (outliers).

Abstract. Sliced inverse regression (SIR) considers a semiparametric regression model between a dependent variable y and a p-dimensional explanatory variable x via a single-index $x'\beta$ and link function f. The direction of the p-dimensional parameter β, called the effective dimension reduction (EDR) direction, is estimated using SIR method. The functional parameter f can be estimated via kernel estimator using the estimated index. However, if outliers are present in the data, this two-step methodology no longer works properly. The aim of this communication is to present three computational methods in order to detect outliers in this single-index model. The numerical behaviors of the proposed outlier detection methods, implemented in R, are illustrated by a simulation study. An example based on a real data is also presented.

Keywords. Semiparametric regression model, Sliced Inverse Regression (SIR), Kernel regression, Outlier detection.

1 Introduction

De nombreux modèles de régression paramétrique (comme le modèle linéaire gaussien) ont été proposés dans la littérature afin d'étudier la relation entre une variable à expliquer $y \in \mathbb{R}$ et une variable explicative p-dimensionnelle $x \in \mathbb{R}^p$. Cependant, trouver la bonne fonction paramétrique de lien entre x et y peut être parfois très complexe. Des modèles non-paramétrique de régression ont alors été proposés, ces modèles sont clairement plus flexibles que les précédents car seules sont faites des hypothèses de régularité sur la fonction de lien entre x et y. Mais, il est bien connu que les estimateurs non-paramétriques (de type estimateurs à noyau par exemple) souffrent du "fléau de la dimension" dès que la dimension p de x devient grande. Afin de contourner ces problèmes des approches purement paramétriques ou purement non-paramétriques, des modèles semi-paramétriques de type "single index" (permettant une réduction de la dimension de la partie explicative du modèle) ont été introduits.

Dans le cadre de la réduction de dimension, la plupart des auteurs supposent que x peut être remplacée par une combinaison linéaire de ses composantes, le fameux index $\beta' x$, sans perte d'information sur la distribution conditionnelle de y sachant x. Une manière d'écrire cette hypothèse est la suivante

$$y \perp x \mid \beta' x \tag{1}$$

où la notation $v_1 \perp v_2 \mid v_3$ signifie que la variable aléatoire v_1 est indépendante de la variable aléatoire v_2 sachant les valeurs prises par la variable aléatoire v_3. On peut également écrire (1) sous la forme, par exemple, d'un modèle de régression semi-paramétrique à un index unique avec un terme d'erreur additif :

$$y = f(\beta' x) + \varepsilon, \tag{2}$$

où f une fonction inconnue à valeurs réelles, la distribution de ε est arbitraire et inconnue, et $\varepsilon \perp x$. Vu que la fonction de lien f est inconnue, le paramètre p-dimensionel β n'est pas totalement identifiable, seul le sous-espace linéaire engendré par β est identifiable. Ce sous-espace est souvent appelé l'espace EDR (pour *effective dimension reduction* en se référant à l'article original de Duan et Li (1991) présentant la méthode SIR (*sliced inverse regression*) qui permet d'estimer une base $b \in \mathbb{R}^p$ de cet espace EDR. On peut noter que cette réduction de dimension peut être une étape très utile dans l'analyse des données disponibles vu que le modèle (1) ne repose que sur très peu d'hypothèses structurelles. Par exemple, il n'est pas nécessaire de prendre un terme d'erreur additif comme dans le modèle (2), ainsi les modèles hétéroscédastiques sont potentiellement inclus dans cette modélisation. De plus, cette réduction de dimension permet de visualiser le lien entre la variable à expliquer et l'indice via le nuage de points croisant les y_i et les indices estimés associés, ce qui fournit une information très utile à la modélisation. Ainsi dans une seconde étape, des approches non-paramétriques standards (tels les splines de lissages ou les estimateurs à noyau) peuvent être mises en oeuvre afin d'estimer la fonction de lien f.

La méthode SIR est connue pour être une méthode pertinente et efficace en réduction de dimension, beaucoup de travaux théoriques ont été faits sur cette méthode et de nombreuses

extensions ont été proposées dans la littérature. Cependant peu d'attention a été portée sur la sensibilité de SIR aux observations atypiques (*outliers*), alors que la théorie de SIR est fondée des propriétés de la matrice de variances-covariances de l'espérance conditionnelle de x sachant y (d'où la présence du terme "régression inverse" dans le nom de la méthode SIR). Il apparaît ainsi évident que la méthode SIR peut être sévèrement influencée par la présence d'observations atypiques dans les données. Quelques méthodes SIR robustes ont été proposées. De même une approche permettant de mettre en évidence les individus influents via les fonctions d'influence a également été développée, mais cette dernière est extrêmement sensible en pratique au choix du nombre H de tranches dans SIR, cet hyperparamètre H permettant de construire l'estimateur de matrice d'intérêt dont une décomposition aux éléments propre dans la méthode SIR. Des références bibliographiques sont disponibles dans Lorenzo et Saracco (2021).

Le but de cette communication est de proposer trois méthodes computationnelles permettant de détecter des *outliers* dans le contexte d'un modèle de type SIR, à savoir un modèle de régression semi-paramétrique à un seul indice, en prenant en compte l'estimation de l'espace EDR via SIR et l'estimation de la fonction de lien f via un estimateur à noyau. Plus précisément, un *outlier* correspond à une observation (x_i, y_i) qui ne proviendrait pas du modèle sous-jacent $y_i = f(\beta' x_i) + \varepsilon_i$. Ainsi un *outlier* est ici une observation ne satisfaisant pas la relation entre y et x définie par le modèle semiparamétrique de régression (1). Un *outlier* (tel que nous l'avons défini) peut clairement ne pas être atypique si l'on se focalise uniquement sur sa valeur x_i ou bien uniquement sur sa valeur y_i. Ainsi, par exemple, ce type d'*outliers* n'est pas détectable en explorant uniquement la distribution des x_i. Notons que s'il y avait présence d'*outliers* visibles dans l'échantillon des x_i, les individus correspondants auraient généralement été détectés dans une étape préliminaire d'étude des données, et le jeu de données aurait alors été "nettoyé" en conséquence avant de faire la modélisation avec un modèle de type SIR.

En pratique, il est toujours intéressant de détecter les *outliers* (plutôt que de développer seulement des méthodes robustes), de les isoler et de comprendre pourquoi ces observations sont atypiques ou aberrantes (mauvaises valeurs numériques ? individus hors normes ? ...). Une fois que les *outliers* ont été identifiés, il convient alors de les supprimer de l'échantillon. Ainsi, sur ces données "nettoyées", il est à nouveau possible de mettre en oeuvre la méthodologie usuelle d'estimation en deux étapes du modèle de régression sous-jacent : SIR pour estimer la direction de β, suivie d'une estimation non-paramétrique de f. Le modèle est ainsi convenablement estimé.

À la section suivante, les méthodes computationnelles de détection des *outliers*, appelées MONO, TTR et BOOT ci-après, seront présentées. Ces méthodes utilisent les erreurs de prédictions IB (*in-bags*) ou OOB (*out-of-bags*) et s'appuient sur des approches de type sous-échantilonnage ou ré-échantillonnage afin de discriminer les *outliers* des observations "normales" (i.e. qui ne sont pas hors normes). Elles ont été implémentées dans R et sont disponibles à l'adresse suivante :

`https://github.com/hlorenzo/outlierSIR`

2 Trois méthodes computationnelles de détection des *outliers* dans la méthode SIR

Considérons l'échantillon $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$ de n individus parmi lesquels se trouvent peut-être des individus atypiques ou *outliers*.

Pour chacune des trois méthodes proposées, le paramètre euclidien β (plus précisément la direction EDR b) est estimé avec la méthode SIR usuelle (avec un nombre de tranches fixé à $H = 10$) et la fonction de lien f est estimée avec un estimateur à noyau (avec un noyau Gaussien et la largeur de fenêtre optimale choisie par Cross-Validation).

2.1 Une approche naïve : la méthode MONO

La méthode MONO repose sur les 3 étapes suivantes.

ETAPE 1. À partir de l'échantillon S,

1.a. Estimation de la direction EDR b. La méthode SIR classique fournit un estimateur $\hat{b}_{\text{SIR}}$ de b. Les indices correspondants $\{\hat{b}'_{\text{SIR}} x_i,\ i = 1, \ldots, n\}$ sont ensuite calculés.

1.b. Estimation des valeurs ajustées $f(\beta' x_i)$ pour $i = 1, \ldots, n$. À partir de l'échantillon $\{(\hat{b}'_{\text{SIR}} x_i, y_i),\ i = 1, \ldots, n$, les valeurs ajustées $\hat{y}_i = \hat{f}_n(\hat{b}'_{\text{SIR}} x_i)$ pour $i = 1, \ldots, n$ sont obtenues via l'estimateur à noyau de la régression $\hat{f}_n(.)$.

ETAPE 2. Évaluation des erreurs associées.

Les erreurs considérées ici sont naturellement les résidus : pour $i = 1, \ldots, n$, $\hat{e}_i = y_i - \hat{y}_i$.

ETAPE 2. Détection des *outliers*.

La détection de potentiels *outliers* est simplement basée sur la définition des *outliers* dans le boxplot des erreurs en valeur absolue $|\hat{e}_i|$ pour $i = 1, \ldots, n$, i.e. les individus dont les valeurs sont plus grandes que les troisième quartile plus 1,5 fois l'écart inter-quartile.

Le nom "MONO" donné à cette méthode reprend le fait que l'échantillon initial S a été utilisé une seule fois, ainsi que le modèle semi-paramétrique sous-jacent.

2.2 La méthode TTR

Cette méthode repose sur la réplication d'échantillons d'apprentissage (*training sample*) et d'échantillons de test (*test sample*) afin d'évaluer la "stabilité" du modèle estimé. Le nom "TTR" de la méthode fait ainsi référence à *Training Test Replications*.

Soit R le nombre de réplications choisi par l'utilisateur. En pratique $R = 2000$ est largement suffisant pour une taille d'échantillon raisonnable, i.e. $n \leq 500$. Soit $\alpha \in [0, 1]$ la proportion de l'échantillon qui va constituer l'échantillon test. Dans la suite, ce paramètre est fixé à $\alpha = 0.1$,

ainsi 90% des individus de l'échantillon initial S sont utilisés pour l'échantillon d'apprentissage S_{train} et les 10% restants constituent l'échantillon test S_{test}. Notons que le tirage des individus est fait à probabilité égale et sans remplacement.

ETAPE 1. Pour chaque réplication r (avec $r = 1, \ldots, R$),

1.a. L'échantillon initial S est partagé en un échantillon d'apprentissage $S^{(r)}_{\text{train}}$ et un échantillon test $S^{(r)}_{\text{test}}$ contenant respectivement $(1-\alpha)\%$ et $\alpha\%$ des individus.

1.b. En utilisant $S^{(r)}_{\text{train}}$, la direction EDR est estimée $\hat{b}^{(r)}_{\text{SIR}}$ et les indices associés $\{(\hat{b}^{(r)}_{\text{SIR}})'x_i,\ i \in S^{(r)}_{\text{train}}\}$ sont calculés.

1.c. Pour tous les individus $i^* \in S^{(r)}_{\text{test}}$, l'erreur de prédiction de la variable à expliquer y est calculée comme suit :

$$\hat{e}^{(r)}_{i^*} = y_{i^*} - \hat{f}^{(r)}_n\left((\hat{b}^{(r)}_{\text{SIR}})'x_{i^*}\right),$$

où l'estimateur à noyau $\hat{f}^{(r)}_n(.)$ est construit sur l'échantillon $\{((\hat{b}^{(r)}_{\text{SIR}})'x_i, y_i),\ i \in S^{(r)}_{\text{train}}\}$.

ETAPE 2. Evaluation des erreurs moyennes.

Pour chaque $i^* = 1, \ldots, n$, la moyenne (sur les R réplications, lorsque l'individu i^* est présent dans l'échantillon test) des erreurs associées est calculée :

$$\overline{e}_{i^*} = \frac{\sum_{r=1}^{R} \left| \hat{e}^{(r)}_{i^*} \right| \mathbb{I}_{[i^* \in S^{(r)}_{\text{test}}]}}{\sum_{r=1}^{R} \mathbb{I}_{[i^* \in S^{(r)}_{\text{test}}]}}.$$

ETAPE 3. Détection des *outliers* via une méthode de détection de rupture.

L'idée est de déterminer un point de rupture unique dans la séquence des erreurs moyennes $\{\overline{e}_{(i^*)},\ i^* = 1, \ldots, n\}$ ordonnées par valeurs décroissantes (où l'indice (i^*) indique qu'il s'agit de la i^*-ème statistique d'ordre de l'échantillon). En effet, s'il n'y a pas *outliers*, aucune rupture ne devrait apparaître clairement dans cette séquence. D'un autre côté, en présence d'*outliers*, les erreurs absolues moyennes correspondantes devraient naturellement être significativement plus grandes que celles associées aux autres individus "normaux". Ainsi, recherche ce point de rupture (en moyenne et en variance) dans la séquence ordonnée des erreurs absolues moyennes devrait intuitivement permettre de séparer les *outliers* des autres observations. Pour cela, le package R `changepoint` a été utilisé (avec l'algorithme de segmentation binaire) pour détecter un unique point de rupture.

Un individu associé à une erreur absolue moyenne (ordonnée) située avant le point de rupture est alors considéré comme un *outlier*.

2.3 La méthode BOOT

La méthode MONO considère les erreur IB (*in-bag*) et la méthode TTR les erreurs OOB (*out-of-bag*). Alors que l'approche MONO peut souffrir de sur-apprentissage (*overfitting*), l'approche TTR risque une perte significative de puissance (vu que l'échantillon d'apprentissage est un sous-échantillon de l'échantillon initial) et ne prend pas en compte l'impact des erreurs IB. La méthode BOOT va ainsi utiliser les erreurs IB dans cet objectif.

Des individus isolés, dans le nuage de points croisant les indices estimés et la variable à expliquer y, qui ne sont pas des *outliers* sont généralement difficiles à prédire, notamment lorsque ces individus ne sont pas dans l'échantillon d'apprentissage. Cependant, si un de ces individus est inclus dans la base de données d'apprentissage, cela aura un effet bénéfique sur le modèle estimé. En effet, ces individus seront alors mieux prédits alors que les individus non isolés seront quant à eux toujours bien prédits puisque ces individus isolés sont en accord avec le modèle de régression. Pour ces individus isolés, l'erreur OBB sera donc importante tandis que l'erreur IB sera potentiellement faible. Ils seront appelées individus *"borderline"* dans la suite. Par contre, les *"outliers"* sont toujours mal prédits avec des erreurs IB et OBB élevées, et les individus "normaux" seront toujours bien prédits avec des faibles erreurs IB et OBB.

La méthode BOOT est basée sur deux règles de décision simples afin de discriminer trois types d'individus (les observations "normales", les observations *"borderline"*, et les *outliers*) en utilisant l'erreur IB et sa transformation logarithmique.

La méthode BOOT repose sur des réplications "bootstrap" de S, d'où son nom. Soit B le nombre d'échantillons "bootstrap" choisi par l'utilisateur. En pratique, $B = 2000$ est largement suffisant pour un échantillon de taille raisonnable, i.e. $n \leq 500$. Notons que les individus sont tirés avec probabilité égale et avec remplacement.

Etape 1. Pour $b = 1, \ldots, B$,

1.a. Un échantillon bootstrap $S^{(b)}$ est tiré à partir de l'échantillon initial S. Soit $n_i^{(b)}$ le nombre de fois où l'observation i est présente dans l'échantillon $S^{(b)}$.

1.b. En utilisant l'échantillon bootstrap $S^{(b)}$, la direction EDR $\hat{b}_{\text{SIR}}^{(b)}$ est estimée et les indices associés $\{(\hat{b}_{\text{SIR}}^{(b)})'x_i,\ i \in S^{(b)}\}$ sont calculés.

1.c. Pour chaque individu $i \in S^{(b)}$, les erreurs IB de prédiction de la variable y sont calculées :

$$\hat{e}_i^{(b)} = y_i - \hat{f}_n^{(b)}\left((\hat{b}_{\text{SIR}}^{(b)})'x_i\right),$$

où l'estimateur $\hat{f}_n^{(b)}(.)$ est construit sur l'échantillon $\{((\hat{b}_{\text{SIR}}^{(b)})'x_i, y_i),\ i \in S^{(b)}\}$.

Etape 2. Evaluation des erreurs moyennes.

Pour chaque $i = 1, \ldots, n$, l'erreur moyenne associée est calculée sur les B réplications (quand l'observation i est présente au moins une fois dans l'échantillon bootstrap corres-

pondant) :

$$\overline{\overline{e}}_{(i)} = \frac{\sum_{b=1}^{B} \left| \hat{e}_i^{(b)} \right| \mathbb{I}_{[i \text{ tel que } n_i^{(b)} \geq 1]}}{\sum_{b=1}^{B} \mathbb{I}_{[i \text{ tel que } n_i^{(b)} \geq 1]}}.$$

ETAPE 3. Détection des *outliers* et des observations "*bordeline*".

L'idée est de tout d'abord identifier parmi les erreurs $\{\overline{\overline{e}}_{(i)},\ i = 1, \ldots, n\}$ celles qui sont particulièrement élevées et qui vont naturellement correspondre aux *"big" outliers*. Pour détecter ces individus fortement atypiques, l'échelle logarithmique est utilisée. Ensuite, dans un second temps, l'échelle d'origine est utilisée afin d'identifier les possibles *"small" outliers* restants, ces observations sont alors dites "borderline".

3.a. La détection des *("big") outliers* potentiels est fondée sur la définition des *outliers* dans le boxplot[1] des $\log\left(\overline{\overline{e}}_{(i)}\right)$ pour $i = 1, \ldots, n$.

3.b. La détection de potentielles observations *"borderline"* est fondée sur la définition des *outliers* dans le boxplot des $\overline{\overline{e}}_{(i)}$ pour $i = 1, \ldots, n$. Les observations "borderline" sont alors définies comme les *outliers* courants qui n'ont pas été identifiées comme des *("big") outliers* à l'étape 3.a. précédente.

Remarque. A l'étape 3.a., la transformation "log" est utilisée par défaut pour détecter les potentiels *outliers*. Cependant, la transformation pertinente à considérer pour les erreurs, $\overline{\overline{e}}_{(i)},\ i = 1, \ldots, n$, n'est probablement pas toujours le logarithme et peut dépendre du modèle (inconnu) de régression sous-jacent (2), en particulier de la fonction de lien f et de la distribution du terme d'erreur ϵ.

3 Remarques finales

Lors de la présentation orale, ces trois approches de détection d'individus atypiques dans SIR seront illustrées sur un exemple simulé et sur un jeu de données réelles. Les comportements numériques des méthodes MONO, TTR et BOOT seront également comparés dans une étude sur simulations. Ces résultats sont décrits dans Lorenzo et Saracco (2021).

Bibliographie

Duan, N., Li, K. C. (1991). Slicing regression: a link-free regression method. *Ann. Stat.*, 19, 505–530.

Lorenzo, H., Saracco, J. (2021). Computational outlier detection methods in sliced inverse regression. Chapter to appear.

[1] déjà décrite dans la présentation de la méthode MONO

A NONPARAMETRIC SPATIAL SCAN STATISTIC FOR FUNCTIONAL DATA

Zaineb SMIDA & Lionel CUCALA & Ali GANNOUN

Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier, France.

E-mail: zaineb.smida@umontpellier.fr ; lionel.cucala@umontpellier.fr ; ali.gannoun@umontpellier.fr

Résumé. Dans ce travail, nous introduisons une méthode non paramétrique de balayage pour des données fonctionnelles indexées dans l'espace. Nous présentons une statistique de balayage construite en utilisant la statistique de test de Wilcoxon-Mann-Whitney pour des données de dimension infinie. Cette dernière est totalement non paramétrique car elle ne suppose aucune distribution concernant les marques fonctionnelles. Ce test de balayage semble puissant contre toute alternative d'agrégation. Nous appliquons cette méthode à un ensemble de données pour extraire des caractéristiques de l'évolution démographique de provinces espagnoles.

Mots-clés. Détection d'Agrégats, Données Fonctionnelles , Espace de Hilbert, Statistique de Balayage Spatiale, Test de Wilcoxon-Mann-Whitney.

Abstract. In this work, we introduce a nonparametric scan method for functional data indexed in space. The scan statistic we present is derived from the Wilcoxon-Mann-Whitney test statistic defined for infinite dimensional data. It is completely nonparametric as it does not assume any distribution concerning the functional marks. This scan test seems to be powerful against any clustering alternative. We apply this method to a data set for extracting features in Spanish province population growth.

Keywords. Cluster Detection, Functional Data, Hilbert Space, Spatial Scan Statistic, Wilcoxon-Mann-Whitney test.

1 Introduction

Cluster detection has become a fruitful area of statistics that has particulary expanded in recent decades. It is used to identify aggregations of events in time and/or space. One of the most popular cluster detection technique is the scan statistic which was firstly introduced by Naus (1963). These scan statistics are used to decide whether exceptional or not observing a cluster of events.
During the last decades, Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed

spatial scan statistics based on Bernouilli and Poisson models. They presented a method based on the likelihood ratio and they tested the clusters' statistical significance via a Monte-Carlo procedure. In the multivariate case, scan statistics based on likelihood ratio were recently tackled by Shen and Jiang (2014) and Cucala et al. (2017). However, in these latter, the likelihood ratio used to construct the scan statistics are computed when the data follow a Gaussian model. A natural question arises: how can we detect a spatial cluster when the data are not Gaussian? In order to overcome this problem, researchers consider the nonparametric procedures which are applicable in many cases where the data are not drawn from a population with a specific distribution.
In the last few years, Jung and Cho (2015) and Cucala (2016) proposed separately a nonparametric spatial scan statistic. In their works, they introduced a scan statistic in the univariate setting which is based on the Wilcoxon-Mann-Whitney test. Very recently, Cucala et al. (2019) proposed a nonparametric scan statistic in the multivariate setting using the Wilcoxon-Mann-Whitney test introduced by Oja and Randles (2004).
Currently, the development of the sensoring allows us to work with huge datasets. Hence, we have more and more access to functional data coming from various fields of applications like environmetrics, medecine and econometrics (see, Ramsay and Silverman (2005), Ferraty and Vieu (2006)).
In the present work, we develop a nonparametric spatial scan statistic for functional data. In Section 2, we explain how the use of the Wilcoxon-Mann-Whitney statistic proposed by Chakraborty and Chaudhuri (2015) can give birth to a scan statistic. Then, to evaluate its statistical significance, we introduce a test procedure based on permutations. In section 3, we apply the spatial scan statistic to simulated and real datasets.

2 Nonparametric spatial scan statistic in functional data

2.1 Statistic construction

Consider X a random element in a separable Hilbert space χ. We denote by $\|.\|_\chi$ a norm on χ. Let $X_1, \ldots, X_n$ be observations of X measured in n different spatial locations $s_1, \ldots, s_n$ included in $D \subset \mathbb{R}^2$. Following the terminology of point process theory, D is the observation domain and X_i is the mark associated to location s_i, for all $i = 1, \ldots, n$. Our goal is to detect a cluster of unusual marks, i.e. a spatial zone $Z \subset D$ in which the marks are abnormally higher or abnormally lower than elsewhere. In order to do that, we will construct a scan statistic which is usually defined to be the maximum of a concentration index observed in a collection of potential clusters using a variable window (see, Nagarwalla (1996)).
In this work, without loss of generality, we consider the circular clusters introduced by Kulldorff (1997). Hence, the set of potential clusters $\mathcal{S}$ is defined as follows:

$$\mathcal{S} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\},$$

where $D_{i,j}$ is the disc centred on s_i and passing through s_j.
Recently, Chakraborty and Chaudhuri (2015) proposed an extension of the Wilcoxon-Mann-Whitney test in the functional case using a spatial sign function defined as $\mathbf{SGN}_x = x/\|x\|_\chi$ for any non zero $x \in \chi$ and $\mathbf{SGN}_x = 0$ if $x = 0$.
Now, we suppose that $X_1, \ldots, X_n$ are independent observations of X (this is a classical assumption in scan statistics). Let $Z \in \mathcal{S}$ be any potential cluster of size n_Z, where $n_Z = \sum_{i=1}^{n} \mathbb{1}(s_i \in Z)$ and Z^c its complement of size $n_{Z^c} = n - n_Z$. Assume that the marks in Z and Z^c respectively follow probability measures P and Q on χ. We suppose that P and Q differ by a shift $\Delta \in \chi$ in the location. For testing the hypothesis $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$, a Wilcoxon-Mann-Whitney test statistic extension in such space is defined as:

$$T_{\mathrm{WMW}} = \frac{1}{n_Z n_{Z^c}} \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \mathbf{SGN}_{\{X_j - X_i\}} = \frac{1}{n_Z n_{Z^c}} \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_i}{\|X_j - X_i\|_\chi}.$$

Chakraborty and Chaudhuri (2015) studied the asymptotic distribution of T_{WMW} and proved the following convergence theorem :
under H_0, if $n_Z/n \to \gamma \in (0,1)$ as $n_Z, n_{Z^c} \to \infty$, then

$$(n_Z n_{Z^c}/n)^{1/2}(T_{\mathrm{WMW}}) \text{ converges weakly to } G(0,\Gamma),$$

where $G(m,C)$ is the distribution of a Gaussian random element in χ with mean $m \in \chi$ and covariance C. Since the covariance function Γ does not depend on n_Z and n_{Z^c}, we can use

$$\mathrm{U}(Z) := (n_Z n_{Z^c}/n)^{1/2} T_{\mathrm{WMW}}$$

as a concentration index to compare potential clusters having different population sizes. Thus, the nonparametric functional scan statistic (NPFSS) is

$$\Lambda_{\mathrm{NPFSS}} = \max_{Z \in \mathcal{S}} \|\mathrm{U}(Z)\|_\chi$$

and the potential cluster detected, for which Λ_{NPFSS} is obtained, is

$$\hat{\mathrm{C}} = \arg\max_{Z \in \mathcal{S}} \|\mathrm{U}(Z)\|_\chi.$$

It is named the most likely cluster (MLC).

2.2 Rule of decision

After computing the scan statistic Λ_{NPFSS} and the most likely cluster $\hat{\mathrm{C}}$, it is necessary to evaluate its significance. However, the distribution, under H_0, of a variable window scan statistic has no analytical form. To overcome this problem, we used a strategy called "random labelling", which was already used in numerous scan methods (see for example, Cucala et al. (2019), Cucala (2017)). This method is based on random permutations and leads to an unbiased estimation of the significance value, whatever the distribution of the data.

3 Application

3.1 Simulation study

In this section, we compared Λ_{NPFSS} with the univariate spatial scan statistic introduced by Cucala (2016), denoted by Λ_{NPUSS}, which is applied to the mean values of the curves. Artificial datasets were generated by using the geographic locations of the 94 french administrative areas named as "*départements*". Each location associated to each "*département*" was defined as its administrative center. The true cluster, denoted by C, is a set of 8 "*départements*" in the Parisian area. We set $\chi = L^2[0,1]$. For all $i \in [1,94]$, the functional marks X_i are generated using the Karhunen-Loève decomposition and they are measured at 101 equispaced points in $[0,1]$. We have considered two different cases: (i) a Gaussian distribution $\mathcal{N}(0,1)$ and (ii) a Student distribution $t(5)$. The probability measures of the marks inside and outside the cluster C differ by a shift $\Delta(t) = \mathbf{c}t, \mathbf{c} \geq 0$ for all $t \in [0,1]$. We generated 100 simulated datasets to see the performance of Λ_{NPFSS} and Λ_{NPUSS} and we computed three distinct criteria: the power to detect a significant cluster, the true positive (TP) rate and the false positive (FP) rate where a type I error equal to 5% was condidered for the rejection of H_0. The following Table 1 gives the results obtained.

		Normal distribution		**Student distribution**	
c		Λ_{NPFSS}	Λ_{NPUSS}	Λ_{NPFSS}	Λ_{NPUSS}
0.0	Power	**0.060**	0.060	0.040	0.040
	%TP	**0.500**	0.500	0.750	0.750
	%FP	**0.475**	0.508	**0.512**	0.689
1.0	Power	**0.210**	0.180	**0.170**	0.150
	%TP	**0.810**	0.799	**0.743**	0.725
	%FP	**0.259**	0.307	**0.276**	0.300
2.0	Power	**0.800**	0.720	**0.580**	0.440
	%TP	**0.975**	0.951	**0.940**	0.920
	%FP	**0.072**	0.078	**0.110**	0.115
3.0	Power	**1.000**	0.980	**0.929**	0.880
	%TP	**0.995**	0.989	**0.977**	0.964
	%FP	**0.021**	0.051	**0.047**	0.065

Table 1: Power, %TP and %FP results of Λ_{NPFSS} and Λ_{NPUSS} when $\Delta(t) = \mathbf{c}t$ in the cases (i) and (ii).

As expected, both methods perform better when the cluster intensity $\mathbf{c}$ becomes larger and our functional scan statistic gives better results since it exploits the whole information contained in the curves (not only the mean values).

3.2 Application to real data

Here, we numerically illustrate how our scan statistic model can be applied to real data. In particular, we considered data for extracting features in Spanish province population growth presented in the study of Cronie et al. (2019).
We considered the demographical evolution of the Spanish province population provided

by the *Spanish Institute of Statistics* (`www.ine.es`). The boundary and centre coordinate data of the 47 provinces of Spain are obtained from the *R* package *raster* .
Our objective here is to detect a spatial area where the demographic evolution would be significantly higher or lower. In order to identify such a cluster, we computed the functional scan statistic on this dataset: $\Lambda_{\text{NPFSS}} = 2.72025$. Based on $T = 999$ permutations, this value is highly significant and $\hat{\text{C}}$ is plotted in Figure 1A. We can see the demographic evolution curves associated to $\hat{\text{C}}$ in the Figure 1B.

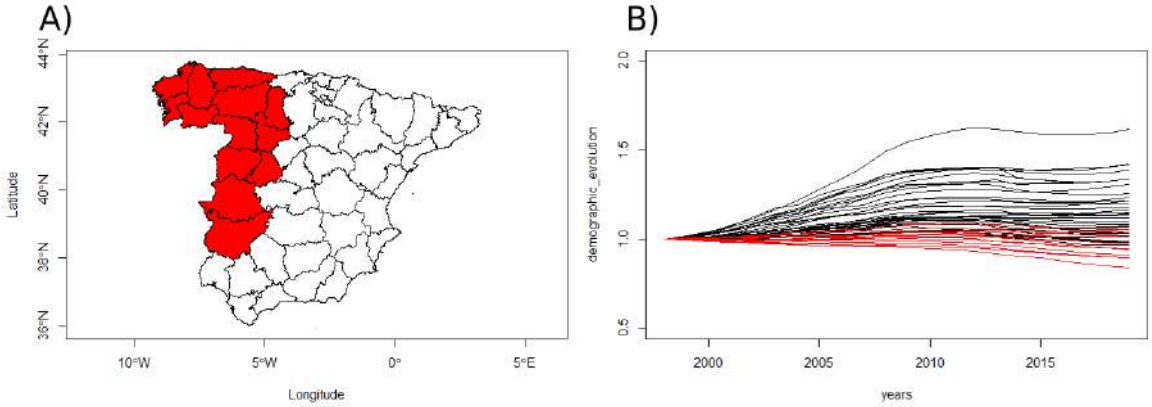


Figure 1: A) : The MLC is presented in red. B) : The demographic evolution curves (from 1998 to 2019) in each province are presented. In red : curves correspond to provinces inside the MLC. In black : curves correspond to provinces outside the MLC.

We remark that the MLC includes 13 locations in the west of Spain (*Asturias*, *Galicia*, *Extremadura* and the west of *Castilla y León*) in which the marks are significantly lower than in the rest of the geographical area studied. We can see that this cluster includes the provinces which have the lowest demographic evolution compared to the other provinces of Spain. This can be explained by a higher mortality rate and a lower birth rate in these regions which have been highly affected by the economic crisis.

4 Conclusion

In this work, we have proposed a nonparametric spatial scan statistic using the Wilcoxon-Mann-Whitney two-sample test for functional data (see, Chakraborty and Chaudhuri(2015)). This scan statistic allows to detect clusters in functional data indexed by space without assuming anything about their distribution.
To do that, we decided to construct a nonparametric spatial scan statistic in the functional case, similar to the one proposed by Cucala (2016) in the univariate case and the one introduced by Cucala et al. (2019) in the multivariate case. First, we proposed a nonparametric scan statistic for functional data in Hilbert space. Second, we defined how to compute its significance using a Monte-Carlo procedure which provides an approximation

to the null distribution. Then, we used artificial and real datasets to see the performance of this scan test.
Recently, Frévent et al. (2020) proposed a parametric spatial scan statistic, denoted by Λ_{PFSS}, which is derived from the functional ANOVA test. In their work, they compared Λ_{NPFSS} with their statistic. They conclude, with simulation studies, that the nonparametric methods performs better against non Gaussian data.

Bibliography

Chakraborty, A. and Chaudhuri, P. (2015). A Wilcoxon-Mann-Whitney type test for infinite-dimensional data. *Biometrika.***102,** 1, 239–246.
Cronie, O., Ghorbani, M., Mateu, J. and Yu, J. (2019). Functional marked point processes–A natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. *arXiv:1911.13142v1 [math.ST].*
Cucala, L. (2016). A Mann-Whitney scan statistic for continuous data. *Communications in Statistics - Theory and Methods.***45,** 321–329.
Cucala, L., Genin, M., Lanier, C. and Occelli, F. (2017). A Multivariate Gaussian scan statistic for spatial data. *Spatial Statistics.* **21,** 66-74.
Cucala, L., Genin, M., Occelli, F. and Soula, J. (2019). A Multivariate nonparametric scan statistic for spatial data. *Spatial Statistics.* **29,** 1-14.
Ferraty, F. and Vieu, Ph. (2006). *Nonparametric Functional Data Analysis (Theory and practice).* Springer-Verlag, New York.
Frévent. C., Ahmed. M.S., Marbac. M. and Genin. M. (2020). Detecting spatial clusters on functional data: a parametric scan statistic approach. *arXiv:2011.03482.*
Jung, I. and Cho, H. (2015). A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics.* **14,** 30.
Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine.* **14,** 799–810.
Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods.* **26,** 1481–1496.
Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in medicine.* **15,** 845–850.
Naus, J. (1963). *Clustering of random points in the line and plane.* Ph.D. Thesis. Rutgers University, New Brunswick, NJ.
Oja, R. and Randles, H.R. (2004). Multivariate nonparametric tests. *Statistical Science.* **19,** 598–605.
Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis (Second edition).* Springer-Verlag New York.
Shen, X. and Jiang, W. (2014). Multivariate normal spatial scan statistic for detecting the most severe cluster of a disease. *Journal of Management Analytics.* **1,** 130–145.

Liste des auteurs