Jeudi 26 Mai 2011

Ce document rassemble les résumés longs des communications.
Pour chaque session, **l'ensemble des résumés courts précède les résumés longs**.

Table des matières

Résumés des conférenciers invités

Jeudi 26 Mai 2011, 8h30-9h15

FALSE DISCOVERY RATE ESTIMATION FOR FREQUENTIST PHARMOCOVIGILANCE SIGNAL DETECTION METHODS

Ismaïl Ahmed,, Inserm U780, Villejuif, France

Pharmacovigilance systems aim at early detection of adverse effects of marketed drugs. They maintain large spontaneous reporting databases for which several automatic signaling methods have been developed. One limit of those methods is that the decision rules for the signal generation are based on arbitrary thresholds. In this article, we propose a new signal-generation procedure. The decision criterion is formulated in terms of a critical region for the P-values resulting from the reporting odds ratio method as well as from the Fisher's exact test. For the latter, we also study the use of mid-P-values. The critical region is defined by the false discovery rate, which can be estimated by adapting the P-values mixture model based procedures to one-sided tests. The methodology is mainly illustrated with the location-based estimator procedure. It is studied through a large simulation study and applied to the French pharmacovigilance database.

Jeudi 26 Mai 2011, 9h15-10h

CONCEPT OF ORTHOGONAL VARIATION IN MUTIL-BLOCK MODELING

Johan Trygg, Umea University

Exactly forty years ago, in 1971, the discipline of chemometrics was born, and has now grown into a well-established multivariate data analysis toolbox in experimental life sciences such as chemistry, biology and medicine. This is in many ways due to the groundbreaking developments during the 80ś of PLS theory and related methods. This laid the foundation for the success and respect our discipline has gained today. During the 90ś, the fields of multivariate calibration, quantitative structure-activity modeling, and multivariate statistical process control were very successful on exploring the predictive capacities of PLS and related regression methods. In biology and medicine however, chemometrics had been largely overlooked in favor of traditional statistics. It was not until around year 2000 and onwards, when the overwhelming size and complexity of the "'omics" technologies (genomics, proteomics, metabolomics), drove biologists and clinicians toward the adoption of chemometrics. Starting around year 2000, biology and medicine forced a shift in focus to model interpretation. In this respect, the concept of Orthogonal variation, introduced in the landmark OSC (1998) and OPLS (2002) papers were in hindsight perfect timing. Since then, biological and medical applications have been a key factor for its strong development and impact. In short, Orthogonal variation in one data matrix (X) is uncorrelated to another data matrix (Y) and include sampling issues, experimental problems, drift, biological variation and non-linearities. It fits very nicely into the framework of the latent variable, which is fundamental in chemometrics.

I will demonstrate the importance of Orthogonal variation in several applications and introduce recent developments with focus on multi-block modeling and data integration. I will also discuss the exciting opportunities with this concept of Orthogonal variation that could have a similar impact and influence to the field of chemometrics as the development of the PLS method once had.

ALGORITHMES PARTICULAIRES POUR L'ESTIMATION DE PARAMETRES STATIQUES DANS LES MODELES A ESPACE D'ETATS

Arnaud Doucet, University of British Columbia

Les modèles a espace d'états non-linéaires non-Gaussiens sont une classe de modèles très populaires pour les séries temporelles. Ils ont trouve de très nombreuses applications en statistiques, économétrie, épidémiologie, ingénierie etc. Toutefois l'inférence dans ces modèles est complexe et nécessite l'utilisation de méthodes de Monte Carlo sophistiquées.

Quand les paramètres du modèle sont spécifiés, les méthodes de filtrage et lissage particulaires se sont imposées comme les méthodes de référence pour l'estimation optimale de l'état. Dans l'immense majorité des applications, il est cependant nécessaire de calibrer le modèle.

Dans ce cadre les méthodes particulaires classiques sont applicables mais extrêmement inefficaces.

Je présenterai ici plusieurs méthodes particulaires originales développées récemment avec mes collaborateurs pour l'inférence Bayésienne et classique, hors ligne et en ligne, de paramètres statiques dans les modèles a espaces d'états.

Ces méthodes seront appliquées à de nombreux exemples apparaissant en écologie, économétrie et ingénierie.

Jeudi 26 Mai 2011, 16h-16h45

PLANS EXPLORATOIRES POUR EXPERIENCES NUMERIQUES

Luc Pronzato, Laboratoire I3S, Université de Nice-Sophia Antipolis-CNRS

Une approche aujourd'hui courante pour planifier une expérience numérique (des simulations sur ordinateur venant remplacer des expérimentations physiques) consiste à répartir n points expérimentaux (chaque point définissant d entrées du code de simulation) de manière à recouvrir le mieux possible un domaine expérimental admissible. On parle alors de "'space-filling designs"' (SFD). Il est communément admis que les plans de ce type ont de bonnes propriétés en terme d'interpolation d'une fonction complexe (le comportement du phénomène simulé) par un modèle simplifié (appelé parfois émulateur). Ils semblent en tout cas satisfaire l'intuition assez naturelle consistant à explorer un peu partout quand on ne sait quasiment rien a priori. L'objectif de cette présentation est de rappeler quelques propriétés de SFD communément utilisés, de proposer de nouveaux critères dont l'optimisation assure un recouvrement adéquat du domaine expérimental et enfin de présenter quelques motivations pour aller au delà des SFD.

ESTIMATING EXTREMAL DEPENDENCE IN TIME SERIES VIA THE EXTREMOGRAM

Thomas Mikosch, University of Copenhagen

The extremogram is a flexible quantitative tool that measures various types of extremal dependence in a stationary time series. In many respects, the extremogram can be viewed as an extreme-value analog of the autocorrelation function (ACF) for a time series. Under mixing conditions, the asymptotic normality of the sample extremogram was derived in Davis and Mikosch (Bernoulli 2009). Unfortunately, the limiting variance

is a difficult quantity to estimate. Instead we employ the stationary bootstrap to the sample extremogram and establish that this resampling procedure provides an asymptotically correct approximation to the central limit theorem. This in turn can be used for constructing credible confidence bounds for the extremogram. The use of the stationary bootstrap for the extremogram is illustrated in several real data examples.

The cross-extremogram measures cross-sectional extremal dependence in multivariate time series. A measure of this dependence, especially the left tail dependence, is of great importance

in the calculation of portfolio risk. We find that devolatilizing the log-returns of an asset is effective in eliminating extremal dependence. On the other hand, cross-extremal dependence between two or more devolatilized log-returns may still remain.

This suggests that extremal dependence between log-returns of two or more assets cannot be explained solely by the volatility in the two series. Following Geman and Chang (2009), we calculate a return time extremogram which measures the waiting time between rare or extreme events in univariate stationary time series. The return time extremogram suggests the existence of extremal clustering in the return times of extreme events for financial assets. The stationary bootstrap can again provide an asymptotically correct approximation to the central limit theorem and can be used for constructing credible confidence bounds for the distribution of return times.

Résumés des communications

JEUDI 26 MAI 2011, 10h20

Données Fonctionnelles

Wavelet-based clustering for mixed-effects functional models, *Madison Giacofci, Sophie Lambert-Lacroix, Franck Picard and Guillemette Marot*

Un nombre croissant de domaines scientifiques collectent de grandes quantités de données comportant beaucoup de mesures répétées pour chaque individu. Ce type de données peut être vu comme une extension des données longitudinales en grande dimension et le cadre naturel de modélisation est alors l'analyse fonctionnelle pour laquelle les unités de base sont les courbes.

Nous proposons une nouvelle procédure de classification de courbes non-supervisée en présence de variabilité inter-individuelle. Nous utilisons pour cela une décomposition en ondelettes des effets fixes et des effets aléatoires assurant que les effets fixes et aléatoires sont dans le même espace fonctionnel. Nous obtenons ainsi, dans le domaine des ondelettes, un modèle linéaire mixte sur lequel on peut appliquer une procédure de classification.

Notre approche se décompose alors en deux étapes. La première est une étape de réduction de dimension basée sur les techniques de seuillage des ondelettes. La taille conséquente des données rend cette étape fondamentale et notre but est de sélectionner les coefficients les plus informatifs pour la classification. Ensuite, une procédure de classification est appliquée sur les coefficients sélectionnés : l'algorithme EM est utilisé pour avoir une estimation des paramètres par maximum de vraisemblance et prédire conjointement les classes des individus et les effets individuels.

Les propriétés de notre procédure sont validées par une étude de simulation approfondie. Nous illustrons ensuite notre méthode sur des données issues de la biologie moléculaires (données omics) comme les données CGH ou les données de spectrométrie de masse.

Notre procédure est disponible dans le package R "curvclust".

Estimation de quantiles conditionnels avec variable explicative fonctionnelle par projection sur un Espace de Hilbert à Noyau Reproduisant, *Yousri Henchiri, Christophe Crambes and Gannoun*

Ce travail a pour objet l'estimation non paramétrique de quantiles conditionnels, la variable explicative X étant à valeurs dans l'espace fonctionnel $L2([0;1])$ des fonctions de carré intégrable sur $[0,1]$ et la variable à expliquer Y étant à valeurs dans $\mathbb{R}$. Nous proposons un prédicteur via la méthode d'apprentissage, Support Vector Machine (SVM), en projetant les observations sur un Espace de Hilbert à Noyau Reproduisant (EHNR) et en utilisant une procédure des moindres carrés itérés pondérés pour résoudre le problème de minimisation pénalisée qui ne possède pas

de solution avec une écriture explicite. Dans cette procédure, nous pouvons fixer les hyperparamètres du modèle par la technique de validation croisée. Nous appliquons la méthodologie proposée sur un jeu de données simulées.

Tests d'homogénéité pour des données fonctionnelles, *Jean Gérard Aghoukeng Jiofack and Guy Martial Nkiet*

Nous proposons deux méthodes pour tester l'égalité des moyennes d'une variable fonc- tionnelle. La première est basée sur la projection de cette variable dans un sous-espace de dimension finie engendré par les fonctions B-splines ou d'ondelettes. La seconde méthode utilise les projections sur les sous-espaces engendrés par des directions principales obtenues de l'analyse en composantes principales fonctionnelle de cette variable. Les lois asymptotiques sous hypothèse nulle de ces tests sont déterminées. Les simulations sont faites pour évaluer la performance des tests ainsi construits en les comparant avec des méthodes existantes.

Réponse fonctionnelle en cinétique : modélisation et prédiction, *François Wahl, Matthieu Canaud, Céline Helbert and Laurent Carraro*

Il arrive souvent que le modèle postulé pour représenter un système physique ou chimique ne convienne que partiellement. C'est le cas du système présenté ici, modèle cinétique de la dépollution des fumées en sortie d'un moteur diesel. Nous montrons comment l'inadéquation partielle du modèle a été levée, en autorisant les paramètres du modèle à varier suivant les entrées, en s'inspirant de la méthodologie des 'modèles à coefficients variables' (VCM). Le modèle ainsi estimé intègre entièrement la spécificité du modèle cinétique tout en l'adaptant localement aux paramètres d'entrée.

It can happen that a model representing a chemical or a physical system works imperfectly and is only partially correct. It is the case of the example in this paper, kinetic model of smoke depollution from the output of diesel engine, and we show how this difficulty has been solved, by letting the parameters of the model vary along the entries, following the methodology of 'varying coefficient models' (VCM). The new model takes into account all the knowledge of the initial kinetic model, while locally adapting it to the entries.

Régression ridge à noyau pour des variables explicatives et d'intérêts fonctionnelles, *Hachem Kadri, Philippe Preux and Emmanuel Duflos*

Dans cet article, on s'intéresse à la régression fonctionnelle dans le cas où les variables explicatives ainsi que les variables d'intérêts sont de dimension infinie et représentées par des fonctions. Nous introduisons une méthode d'estimation non-paramétrique de la fonction de régression, extension de la régression ridge à noyau au domaine de l'analyse des données fonctionnelles, basée sur la généralisation de la théorie des espaces de Hilbert à noyaux reproduisants (EHNR). L'originalité du travail réside essentiellement dans le choix du noyau reproduisant construit à partir de l'opérateur intégrale et la proposition d'une procédure pour l'inversion de la matrice noyau à blocs opérateurs.

Wavelet-based clustering for mixed-effects functional models

Madison Giacofci[1], Sophie Lambert-Lacroix[2], Guillemette Marot[3,4] & Franck Picard[3]

[1] *Laboratoire LJK, Université de Grenoble et CNRS, UMR 5224, 38041 Grenoble cedex 9, France*

[2] *UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG UMR 5525, Grenoble, F-38041, France*

[3] *Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558 - Univ. Lyon 1, F-69622, Villeurbanne, France*

[4] *Projet BAMBOO, INRIA Rhône-Alpes, F-38330 Montbonnot Saint-Martin, France.*

Résumé

Un nombre croissant de domaines scientifiques s'intéressent aux données comportant beaucoup de mesures répétées pour chaque individu. Ce type de données peut être vu comme une extension des données longitudinales en grande dimension et le cadre naturel de modélisation est alors l'analyse fonctionnelle pour laquelle les unités de base sont les courbes. Notre objectif est de réaliser une classification non supervisée de ces courbes en présence de variabilité inter-individuelle. Les approches existantes sont fondées sur les splines (James et Sugar (2003)). Cependant, ces modèles ne permettent pas de prendre en compte des fonctions présentant des irrégularités et leur utilisation est limitée à des données de faible dimension.

Nous proposons une nouvelle procédure de classification de courbes non-supervisée fondée sur une décomposition en ondelettes des signaux. Nous introduisons un effet aléatoire prenant en compte la variabilité inter-individuelle et grâce à une modélisation appropriée dans le domaine des ondelettes, nous nous assurons que les effets fixes et aléatoires appartiennent au même espace fonctionnel (espace de Besov, Antoniadis et Sapatinas (2007)). Ainsi nous obtenons un modèle de mélange Gaussien multivarié dont les composantes s'écrivent comme des modèles linéaires mixtes.

Nous proposons une procédure en deux étapes. Nous commenons par une étape de réduction de dimension basée sur les techniques de seuillage des ondelettes. La taille conséquente des données rend cette étape fondamentale et notre but est de sélectionner les coefficients les plus informatifs pour la classification. Ensuite, une procédure de classification est appliquée sur les coefficients sélectionnés : l'algorithme EM est utilisé pour avoir une estimation des paramètres par maximum de vraisemblance et prédire conjointement les classes des individus et les effets individuels.

Les propriétés de notre procédure sont validées par une étude de simulation approfondie. Nous illustrons ensuite notre méthode sur des données issues de la biologie moléculaire (données omics) comme les données CGH ou les données de spectrométrie de masse. Notre procédure est disponible dans le package R `curvclust`.

Mots-clés : données fonctionnelles, ondelettes, modèle linéaire mixte, classification non-supervisée, algorithme EM.

Abstract

More and more scientific studies yield to the collection of large amounts of data that consist of sets of curves recorded on individuals. These data can be seen as an extension of longitudinal data in high dimension and are often modeled as functional data. Our purpose is to perform unsupervised clustering of these curves in the presence of inter-individual variability. Curve clustering is a widely studied subject and splines have been proposed to account for inter-individual variability in this context James and Sugar (2003). However splines are known to be computationally inefficient and they can not be used to model irregular curves such as peak-like data.

We develop a new procedure to perform clustering of functional data in the presence of inter-individual variability. We use a wavelet decomposition of the data for both fixed and random-effects. This ensures that both fixed and random effects lie in the same functional space even when dealing with irregular functions that belong to Besov spaces (Antoniadis and Sapatinas (2007)). In the wavelet basis the model resumes to a linear mixed-effects model that can be used for a model-based clustering algorithm.

Our approach follows two steps. First an efficient dimension reduction step based on wavelet thresholding is performed. This first step is necessary due to the high dimensionality of the data. Our aim is to select the wavelet coefficients that are the most informative with respect to the clustering objective of the procedure. Then a clustering step is applied on the selected coefficients. An EM-algorithm is used for maximum likelihood estimation and to predict jointly label variables and random effects.

The properties of the overall procedure are validated by an extensive simulation study. Then we illustrate our method on high throughput molecular data (omics-data) like microarray CGH or mass spectrometry data. Our procedure is available through the R package `curvclust`.

Key-words : functional data, wavelets, linear mixed model, clustering, EM-algorithm.

Bibliographie

[1] Antoniadis A., Sapatinas T.,
Estimation and inference in functional mixed-effects models, *Computational Statistics & Data Analysis, Volume 51, Issue 10, 15 June 2007, Pages 4793-4813*
[2] James, G. and Sugar, C.,
Clustering for sparsely sampled functional data, *Journal of the American Statistical Association, Volume 98, Number 462, June 2003*

Estimation de quantiles conditionnels avec variable explicative fonctionnelle par projection sur un Espace de Hilbert à Noyau Reproduisant.

Yousri HENCHIRI & Christophe CRAMBES & Ali GANNOUN

I3M Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex, France.

Résumé. Ce travail a pour objet l'estimation non paramétrique de quantiles conditionnels, la variable explicative X étant à valeurs dans l'espace fonctionnel $\mathbf{L}^2([0,1])$ des fonctions de carré intégrable sur [0,1] et la variable à expliquer Y étant à valeurs dans $\mathbb{R}$. Nous proposons un prédicteur via la méthode d'apprentissage, *Support Vector Machine (SVM)*, en projetant les observations sur un Espace de Hilbert à Noyau Reproduisant (EHNR) et en utilisant une procédure des moindres carrés itérés pondérés pour résoudre le problème de minimisation pénalisée qui ne possède pas de solution avec une écriture explicite. Dans cette procédure, nous pouvons fixer les hyperparamètres du modèle par la technique de validation croisée. Nous appliquons la méthodologie proposée sur un jeu de données simulées.
Mots clés. *Variable explicative fonctionnelle, Quantiles conditionnels, Support Vector Machine, Problème mal posé, Espace de Hilbert à Noyau Reproduisant.*

Abstract. This work deals with conditional quantiles estimation via non parametric methodology when the explanatory variable X takes values in the functional space $\mathbf{L}^2([0,1])$ of square integrable functions defined on [0,1], and a scalar response Y. We present Support Vector Machine Quantile Regression (SVMQR) by projection of the observations on a Reproducing Kernel Hilbert Space (RKHS) using iterative reweighted least squares (IRWLS) procedure to solve the penalized minimisation problem which has no explicit solution. This procedure makes it possible to derive the cross validation method for choosing the hyper-parameters of the model. The performances of our algorithm, in terms accuracy of the prediction, are evaluated with a simulation study.
Keywords. *Functional covariate, Conditional quantiles, Support Vector Machine, Ill-conditioned inverse problem, Reproducing Kernel Hilbert Spaces.*

1 Introduction

Nous avons assisté ces dernières années à l'explosion de la masse de données disponibles par le développement des moyens de communication où de nombreuses machines informatiques permettent de fournir une information volumineuse et complexe. De telles données, collectées par exemple de façon très fine dans le temps, sont maintenant bien connues sous le nom de données fonctionnelles. Ce domaine fait l'objet de nombreux travaux dans la

communauté statistique *[Ramsay & Silverman (2005)*, *Ferraty & Vieu (2006)* et plus récemment *Ferraty & Romain (2010)]*.

Les quantiles conditionnels sont fréquemment utilisés en statistique, par exemple pour la construction d'intervalles de prédiction, la détermination de courbes de référence ou comme outil de prévision alternatif à la moyenne conditionnelle. La médiane est un indicateur robuste de la tendance centrale d'une population, l'intervalle interquartile est un bon indicateur de la dispersion. Dans la pratique, les domaines d'utilisation des quantiles sont assez variés : économétrie, finance, biologie et environnement. Pour une large revue des méthodes autour des quantiles conditionnels, le lecteur peut se rapporter à *Koenker (2005)* et plus particulièrement à *Christmann & Steinwart (2008)* pour l'application de la méthode *Support Vector Machine* ***(SVM)***.

Dans ce travail, nous allons expliciter la théorie d'apprentissage statistique à travers la **SVM** *[Vapnik (1998)*, *Schölkopf & Smola (2002)* et plus récemment *Steinwart & Christmann (2008)]*. Le critère d'apprentissage est basé sur un *risque régularisé* pour lequel *la complexité du modèle* est mesurée par une norme dans *un Espace de Hilbert à Noyau Reproduisant* ***(EHNR)*** *[Berlinet & Thomas-Agnan (2004)* et plus récemment *Steinwart & Christmann (2008)]*. Des travaux pour le traitement de données fonctionnelles ont déjà été proposés, par exemple : *[Rossi & Villa (2006)] (classification binaire)* et *[Preda (2007)] (régression)*. Nous nous intéressons dans ce travail, via la projection sur un **EHNR**, au problème de l'estimation de quantiles conditionnels lorsque la variable explicative est fonctionnelle et la variable à expliquer est scalaire *[Cardot & al. (2005)]*.

2 Modèle et construction de l'estimateur

Soit un échantillon $D := (X_i, Y_i)_{i=1,\dots,n} \in (\mathcal{X} \times \mathcal{Y})^n$ constitué de $\boldsymbol{n}$ couples de variables aléatoires définies sur un même espace de probabilité, de même loi que *(X, Y)*, où X est à valeurs dans l'espace des entrées $\mathcal{X} := \mathbf{L}^2([0,1])$ des fonctions de carré intégrable sur un intervalle [0,1] et Y appartient à l'espace des sorties $\mathcal{Y}$ à valeurs dans un intervalle ζ de $\mathbb{R}$.

Soit $\tau \in]0,1[$ fixé, pour tout $x \in \mathcal{X}$, le quantile conditionnel d'ordre τ de Y sachant $[X = x]$, noté $q_\tau(x)$, est défini par la relation

$$\mathbf{P}(Y \leq q_\tau(x) | X = x) = \tau. \tag{1}$$

Une condition qui garantit l'existence et l'unicité de $q_\tau(x)$ est de supposer par exemple que la fonction de répartition de Y sachant $[X = x]$ est continue et strictement croissante de ζ vers [0,1].

Le quantile conditionnel est aussi solution du problème de minimisation suivant :

$$\min_{a \in \mathbb{R}} \left\{ \mathbb{E}(\rho_\tau(Y - a) | X = x) \right\}, \tag{2}$$

où $\rho_\tau(r) := |r| \ + \ (2\tau - 1)\ r$ *[Koenker (2005)]*.

Du fait que $X \in \mathcal{X}$ de dimension infinie, le problème de minimisation ci-dessus est mal adapté. On définit alors une estimation du quantile conditionnel comme solution du problème de minimisation :

$$\min_{\Psi \in \mathcal{H}_\mathcal{K}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \Psi(X_i)) \ + \ \lambda \ \|\Psi\|^2_{\mathcal{H}_\mathcal{K}} \right\}, \tag{3}$$

où $\mathcal{H}_\mathcal{K}$ est un **EHNR** d'un noyau reproduisant mesurable $\mathcal{K}$ et λ est un paramètre de pénalisation positif dont le but est de contrôler le degré de régularité de l'estimateur cherché.

Remarques:

[i] Si dans (2), $\rho_\tau(r) := r^2$, le problème de minimisation a pour solution α vérifiant le système linéaire suivant :

$$(\lambda n \mathbf{I} + [\mathcal{K}])\ \alpha = Y, \tag{4}$$

où $\mathbf{I}$ est la matrice identité d'ordre n et $[\mathcal{K}]$ est la matrice de Gram associée au noyau $\mathcal{K}$ avec $[\mathcal{K}]_{i,j} := (\mathcal{K}(X_i, X_j))_{i,j}$, $1 \leq i, j \leq n$.

[ii] Le problème de minimisation (3) s'écrit:

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\rho_\tau(Y_i - \sum_{j=1}^{n} \alpha_j \mathcal{K}(X_i, X_j))) \ + \ \lambda\ \alpha^T\ [\mathcal{K}]\ \alpha \right\}, \tag{5}$$

selon le théorème de représentation *Kimeldorf & Wahba (1971)*, où $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ et :

$$\Psi(x) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, X_i), \quad \forall x \in \mathcal{X}.$$

3 Algorithme de résolution

Nous utilisons, dans notre méthodologie, *l'algorithme des moindres carrés itérés pondérés [Ruppert & Caroll (1988)]*. Le problème de minimisation (5) s'écrit :

$$\min_{\alpha \in \mathbb{R}^n} \left\{ C \ \sum_{i=1}^{n} \Gamma_i(\tau) \left| Y_i - \sum_{j=1}^{n} \alpha_j \mathcal{K}(X_i, X_j) \right| \ + \frac{1}{2}\ \alpha^T\ [\mathcal{K}]\ \alpha \right\}, \tag{6}$$

avec : $\Gamma_i(\tau) = 2\tau\ \mathbf{1}_{\left\{(Y_i - \sum_{j=1}^n\ \alpha_j \mathcal{K}(X_i,X_j)) \geq 0\right\}} + 2(1-\tau)\ \mathbf{1}_{\left\{(Y_i - \sum_{j=1}^n\ \alpha_j \mathcal{K}(X_i,X_j)) < 0\right\}}$ où $\mathbf{1}$ désigne la fonction indicatrice d'un ensemble. Le paramètre C est égal à $\frac{1}{2n\lambda}$.

Le principe de l'algorithme des moindres carrés itérés pondérés consiste à remplacer la valeur absolue par un terme quadratique pondéré. Nous obtenons ainsi, à chaque étape de l'algorithme, une expression explicite de la solution du problème de minimisation.

◦ ***Initialisation***: On détermine $\hat{\alpha}^1$ solution du problème de minimisation

$$\min_{\alpha \in \mathbb{R}^n} \left\{ C\ \sum_{i=1}^n (Y_i - \sum_{j=1}^n\ \alpha_j \mathcal{K}(X_i,X_j))^2\ + \frac{1}{2}\ \alpha^T\ [\mathcal{K}]\ \alpha \right\}. \tag{7}$$

La solution explicite de ce problème est donnée par la relation (4).

◦ ***Étape (q+1)***: Connaissant $\hat{\alpha}^{(q)}$, on détermine $\hat{\alpha}^{(q+1)}$ solution du problème de minimisation

$$\min_{\alpha \in \mathbb{R}^n} \left\{ C\ \sum_{i=1}^n \left(\frac{\Gamma_i^{(q)}(\tau)(Y_i - \sum_{j=1}^n\ \alpha_j \mathcal{K}(X_i,X_j))^2}{[(Y_i - \sum_{j=1}^n\ \alpha_j \mathcal{K}(X_i,X_j))^2 + \Theta^2]^{\frac{1}{2}}} \right)\ + \frac{1}{2}\ \alpha^T\ [\mathcal{K}]\ \alpha \right\}, \tag{8}$$

où Θ^2 est une constante strictement positive que l'on se fixe pour éviter un dénominateur nul et $\Gamma_i^{(q)}(\tau)$ est $\Gamma_i(\tau)$ à l'étape *(q)* de l'algorithme, à savoir
$\Gamma_i^{(q)}(\tau) = 2\tau\ \mathbf{1}_{\left\{Y_i - \sum_{j=1}^n\ \alpha_j^{(q)} \mathcal{K}(X_i,X_j) \geq 0\right\}} + 2(1-\tau)\ \mathbf{1}_{\left\{Y_i - \sum_{j=1}^n\ \alpha_j^{(q)} \mathcal{K}(X_i,X_j) < 0\right\}}$.

En définissant la matrice diagonale $\mathbb{M}^{(q)}$ de taille $n \times n$, dont les éléments diagonaux sont donnés, pour tout $i = 1, \ldots, n$, par :

$$\left(\frac{\Gamma_i^{(q)}(\tau)}{[(Y_i - \sum_{j=1}^n\ \alpha_j^{(q)} \mathcal{K}(X_i,X_j))_i^2 + \Theta^2]^{\frac{1}{2}}} \right),$$

on obtient la solution du problème de minimisation de l'étape *(q+1)* comme suit :

$$\alpha^{(q+1)} = [C^{-1}\ \mathbf{I} + \mathbb{M}^{(q)}[\mathcal{K}]]^{-1}\ \mathbb{M}^{(q)} = [n\ \lambda\ \mathbf{I} + \mathbb{M}^{(q)}[\mathcal{K}]]^{-1}\ \mathbb{M}^{(q)} Y.$$

◦ ***Critère d'arrêt***: On décide d'arrêter l'algorithme lorsque $\left\|\hat{\alpha}^{(q+1)} - \hat{\alpha}^{(q)}\right\|_{\mathbb{R}^n} < err$, où la quantité *err* est fixée.

4 Résultats expérimentaux et sélection des paramètres

La procédure d'estimation exposée précédemment dépend de beaucoup de paramètres : le paramètre de lissage λ (autrement dit le paramètre C) et les paramètres du noyau

utilisé. Ces paramètres jouent un rôle crucial pour donner de bonnes prédictions. Dans notre démarche, nous allons fixer le paramètre λ et les paramètres du noyau par validation croisée *[Wahba (1990)]*. Nous précisons que lors des simulations les courbes X sont discrétisées en $\boldsymbol{p}$ points.

Pour accroître l'efficacité, nous utilisons d'une part des séquences de croissance exponentielle de λ, $\lambda = e^{(-21)}, e^{(-18)}, \ldots, e^{(3)}$, et de h *(la fenêtre du noyau gaussien)*, avec $h = e^{(-10)}, e^{(-8)}, \ldots, e^{(2)}$, et d'autre part des séquences de d *(degré du noyau polynomial)* avec $d = 2, 3, 4, 5$.

Un exemple traité ainsi que les différents paramètres sont précisés ci-dessous :

• $X := \{X_t\}_{t\in[0,1]}$ est un mouvement Browien standard et Y est une variable aléatoire de moyenne nulle définie par

$$Y = \int_0^1 (t - 2/3)\, X_t^2 dt + \epsilon, \quad \epsilon \textit{ est un bruit gaussien } [\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)].$$

- Les noyaux utilisés sont les suivants :
  - Noyau gaussien : $\mathcal{K}(X_1, X_2) = exp(-h \left\|X_1 - X_2\right\|^2_{\mathbf{L}^2([0,1])}), \quad h > 0.$
  - Noyau polynomial : $\mathcal{K}(X_1, X_2) = (1 + \ \prec X_1, X_2 \succ_{\mathbf{L}^2([0,1])})^d, \quad d \in \mathbb{N}^*.$

• $\tau \in]0, 1[$ est l'ordre du quantile conditionnel.

Nous explicitons les résultats pour l'ordre $\tau = 0.5$ *(Table 1)*. Les échantillons ont été simulés $\boldsymbol{S}$ fois *($\boldsymbol{S}$ = 100)* avec différentes tailles ($\mathbf{n}$ = 50, 100, 300) en $\mathbf{p}$ = 100 points de discrétisation. Chaque échantillon a été découpé en un échantillon d'apprentissage *(trois quarts de chaque échantillon simulé)* et en un échantillon de validation *(un quart de chaque échantillon simulé)*.

Le critère utilisé dans la comparaison décrit dans la ***Table 1*** est la moyenne d'Erreur Quadratique Moyenne ($\overline{EQM}$) sur l'ensemble des échantillons de validation simulés :

$$\overline{EQM} = \frac{1}{\boldsymbol{S}} \sum_{i=1}^{\boldsymbol{S}} EQM(i),$$

où $EQM(i) = \frac{1}{n} \sum_{j=1}^{\boldsymbol{n}} (Y_j - \hat{Y}_j)^2$ est l'Erreur Quadratique Moyenne sur le $i^{ème}$ échantillon de validation simulé avec $i \in \{1, \ldots, \boldsymbol{S}\}$.

Bibliographie

[1] Berlinet, A. & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publisher.

[2] Cardot, H., Crambes, C. & Sarda, P. (2005). Quantile regression when the covariates are functions, *Journal of Nonparametric Statistics*, **17**, 841–856.

[3] Christmann, A. & Steinwart, I. (2008). Consistency of kernel based quantile regression, *Applied Stochastic Models in Business and Industry*, **24(2)**, 171–183.
[4] Ferraty, F. & Vieu, P. (2006). *Nonparametric functional data analysis. Theory and practice*. Springer-Verlag, New York.
[5] Ferraty, F. & Romain, Y. (2010). *Handbook on Functional Data Analysis and related fields*. Oxford University Press.
[6] Kimeldorf, G. S. & Wahba, G. (1971). Some results on Tchebycheffian spline functions, *Journal of Mathematical Analysis and Applications*, **33**, 82–95.
[7] Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, Cambridge.
[8] Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods, *Journal of Statistical Planning and Inference*, **137(3)**, 829–840.
[9] Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis (Second edition)*, Springer-Verlag, New York.
[10] Rossi, F. & Villa, N. (2006). Support Vector Machine for Functional Data Classification, *Neurocomputing*, **69(7-9)**, 730–742.
[11] Ruppert, D. & Caroll, J. (1988). *Transformation and weighting in regression*, Chapman and Hall, New York.
[12] Schölkopf, B. & Smola, A. J. (2002). *Learning with Kernels*, MIT Press, Cambridge.
[13] Steinwart, I. & Christmann, A. (2008). *Support Vector Machines*, Springer, New York.
[15] Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley and Sons.
[16] Wahba, G. (1990). *Spline Models for Observational Data*, Series in Applied Mathematics **59**, SIAM, Philadelphia.

Table 1: *Résultat d'analyse.*

Prédiction sur les données de validation ($\overline{EQM}$ $\times 10^{-4}$).

	$\mathbf{n} = 50$	$\mathbf{n} = 100$	$\mathbf{n} = 300$
Données Brutes + SVM (gaussien)	*10.42612*	*9.504034*	*6.03263*
	(5.943963)	*(4.438303)*	*(1.465756)*
Données Brutes + SVM (polynomial)	*9.65324*	*8.12378*	*5.965548*
	(9.338067)	*(3.627916)*	*(1.651072)*
B-spline + SVM (gaussien)	*8.046232*	*6.301614*	*4.715405*
	(5.615603)	*(3.127043)*	*(1.177536)*
B-spline + SVM (polynomial)	*8.203936*	***5.658505***	*4.872072*
	(6.380718)	***(1.88807)***	*(1.118844)*
Base de Fourier + SVM (gaussien)	*7.422663*	*6.143281*	***4.585847***
	(4.154852)	*(2.622473)*	***(0.9638604)***
Base de Fourier + SVM (polynomial)	***6.816327***	*6.440656*	*4.739825*
	(4.422176)	*(2.975028)*	*(1.135096)*

Tests d'homogenéité pour des données fonctionnelles

Jean Gérard AGHOUKENG JIOFACK

Université de Yaoundé I, Faculté des sciences, Département de Mathématiques, BP 812 Yaoundé, Cameroun.
E-mail : aghoukeng-maths@yahoo.fr

Guy Martial NKIET

Université des Sciences et Techniques de Masuku, Faculté des Sciences, Département de Mathématiques et Informatique, BP 943 Franceville, Gabon.
E-mail : gnkiet@hotmail.com

Résumé

Nous proposons deux méthodes pour tester l'égalité des moyennes d'une variable fonctionnelle. La première est basée sur la projection de cette variable dans un sous-espace de dimension finie engendré par les fonctions B-splines ou d'ondelettes. La seconde méthode utilise les projections sur les sous-espaces engendrés par des directions principales obtenues de l'analyse en composantes principales fonctionnelle de cette variable. Les lois asymptotiques sous hypothèse nulle de ces tests sont déterminées. Les simulations sont faites pour évaluer la performance des tests ainsi construits en les comparant avec des méthodes existantes.

Abstract

We propose two methods for testing for equality of means of a functional variable. The first one is based on a projection of this variable onto a finite dimensional subspace spanned by B-spline functions or wavelets. The second method uses projection onto a subspace spanned by principal directions obtained from the functional principal component analysis of the variable. The asymptotic distributions of the related test statistics are obtained under the null hypothesis. Simulations that permit to evaluate the performance of the proposed tests with comparisons with existing methods are given.

Mots-clés : test d'homogenéité ; données fonctionnelles ; Analyse discriminante ; B-splines ; ondelettes.

1 Introduction et position du problème

Soit X une variable aléatoire définie sur (Ω, A, P) et à valeurs dans un espace de Hilbert réel séparable H muni du produit scalaire $\langle \cdot, \cdot \rangle$ et de norme associée $\|\cdot\|$. On suppose que $\mathbb{E}\left(\|X\|^4\right) < +\infty$ et on considère, par ailleurs, une variable aléatoire Y à valeurs dans $F := \{1, ..., q\}$. Pour $\ell \in F$, posant $m_\ell = \mathbb{E}(X|Y = \ell)$ et $p_\ell = P(Y = \ell)$, on suppose sans perte de généralité que $p_\ell > 0$ pour tout $\ell \in F$. On s'intéresse au test d'homogenéité pour la variable aléatoire X, c'est à dire au test d'hypothèse nulle

$$\mathcal{H}_0 \, : \, m_1 = m_2 = ... = m_q$$

contre

$$\mathcal{H}_1 \, : \, \exists \ell \in F, \tau_\ell \neq 0, \text{ où } \tau_\ell = m_\ell - m \text{ et } m = \mathbb{E}(X).$$

Nous introduisons deux méthodes pour ce problème de test. La première est basée sur l'analyse discriminante (AD) de la projection de X sur un sous-espace d'approximation de dimension finie. La deuxième approche est basée sur des projections des $\tau_\ell \, (\ell = 1, ..., q)$ sur un nombre fini d'axes principaux issus de l'ACP de X. Dans chacun des deux cas, une statistique de test, basée sur un échantillon i.i.d. $\left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{1 \leq i \leq n}$ de (X, Y), est définie et sa loi limite sous hypothèse nulle est déterminée.

2 Approche par projection sur un sous-espace de dimension finie

Soit $(E^r)_{r \in \mathbb{N}^*}$ une suite de sous-espaces de dimensions finies de H telle que $E^r \subset E^{r+1}$, pour tout $r \in \mathbb{N}^*$, et $H = \overline{\bigcup_{r \in \mathbb{N}^*} E^r}$. Désignant par $\mathbb{P}^r$ la projection orthogonale de H sur E^r, il est bien connu que pour tout $x \in H$, on a $\lim_{r \to +\infty} \mathbb{P}^r x = x$. Alors, pour r suffisamment grand, puisque $\mathcal{H}_0$ implique que $\mathbb{P}m_1 = \mathbb{P}m_2 = ... = \mathbb{P}m_q$ où $\mathbb{P} := \mathbb{P}^r$, on va considérer un test de cette dernière en nous basant sur l'analyse discriminante (AD) de $\widetilde{X} = \mathbb{P}X$. Posant $E := E^r$, on considère une base $\mathcal{B} = \{\phi_1, ...\phi_p\}$ de E; celle-ci induit l'application linéaire $\mathbb{L} : \alpha \in \mathbb{R}^p \to \sum_{i=1}^{p} \alpha_i \phi_i \in H$. Il est alors facile de vérifier que l'AD précédente est équivalente à celle du vecteur aléatoire $\mathbb{L}^* X$. Celle-ci (cf., par exemple, Dauxois et Nkiet (1997)) est obtenue grâce à l'analyse spectrale de l'opérateur

$$R(p) = \left(W_1^{1/2}\right)^\dagger B_* \left(W_1^{1/2}\right)^\dagger,$$

où $\dagger$ désigne l'inverse généralisée de Moore-Penrose, $B_* = \sum_{\ell=1}^{q} p_\ell \tau_{*\ell} \otimes \tau_{*\ell}$ avec $\tau_{*\ell} = \mathbb{L}^* m_\ell - \mathbb{L}^* m$, et W_1 est l'opérateur de covariance de $\mathbb{L}^* X$. Il est facile de vérifier que si $\mathcal{H}_0$ est

vraie, on a

$$T(p) = tr\left(R(p)\right) = \sum_{\ell=1}^{q} p_\ell \left\| \left(\left(W_1^{1/2}\right)^{\dagger} \left(\mathbb{L}^* m_\ell - \mathbb{L}^* m\right) \right) \right\|_{\mathbb{R}^p}^2 = 0;$$

on peut, par conséquent, utiliser un estimateur de $T\left(p\right)$ comme statistique de test. Soient

$$B_*^{(n)} = \sum_{\ell=1}^{q} p_\ell^n \left(\mathbb{L}^* \overline{X}_\ell^n - \mathbb{L}^* \overline{X}^n\right) \otimes \left(\mathbb{L}^* \overline{X}_\ell^n - \mathbb{L}^* \overline{X}^n\right),$$

$$W_1^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{L}^* X^{(i)} \otimes \mathbb{L}^* X^{(i)} - \mathbb{L}^* \overline{X}^n \otimes \mathbb{L}^* \overline{X}^n$$

où

$$p_\ell^n = \frac{n_\ell}{n}, n_\ell = \sum_{i=1}^{n} \mathbf{1}_{\left\{Y^{(i)}=\ell\right\}}, \overline{X}_\ell^n = \frac{1}{n_\ell} \sum_{i=1}^{n} \mathbf{1}_{\left\{Y^{(i)}=\ell\right\}} X^{(i)}, \overline{X}^n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)},$$

on estime $R\left(p\right)$ par l'opérateur aléatoire $\widehat{R}^{(n)}(p) = W_1^{(n)\dagger} B_*^{(n)}$ et on définit la variable aléatoire $\widehat{T}^{(n)}\left(p\right) = tr\left(\widehat{R}^{(n)}\left(p\right)\right)$ comme statistique de test. Pour déterminer sa loi sous $\mathcal{H}_0$, on suppose que $rg\left(W_1^{(n)}\right)$ converge presque sûrement vers $rg\left(W_1\right)$ lorsque $n \to +\infty$; ceci assure la convergence de $\widehat{R}^{(n)}\left(p\right)$ vers $R(p)$. Soit $\Delta = diag\left(p_1, p_2, ..., p_q\right)$ et $\Gamma_1 = \Delta^{-1} \otimes^K W_1^{\dagger}$, où $\otimes^K$designe le produit de Kronecker; on considère la matrice

$$\Sigma_1 = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & ... & \Sigma_{1q} \\ \Sigma_{21} & \Sigma_{22} & ... & \Sigma_{2q} \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{q1} & \Sigma_{q2} & ... & \Sigma_{qq} \end{pmatrix},$$

où $\Sigma_{ij} = \delta_{ij} p_i V_*^{(i)} + p_i p_j \left(W_1 - V_*^{(i)} - V_*^{(j)}\right)$, δ_{ij} étant le symbole de Kronecker et

$$V_*^{(\ell)} = \mathbb{E}\left[\left(\mathbb{L}^* X - \mathbb{L}^* m_\ell\right) \otimes \left(\mathbb{L}^* X - \mathbb{L}^* m_\ell\right) | Y = \ell\right], \ \ \ell = 1, 2, ..., q.$$

On obtient alors :

Théorème 1. *Sous $\mathcal{H}_0$, $n\widehat{T}_1^{(n)}\left(p\right)$ converge en loi, lorsque $n \to +\infty$, vers $Q_1 = M_1^T \Gamma_1 M_1$ où M_1 est un vecteur aléatoire gaussien dans $\mathbb{R}^{pq}$, centré et d'opérateur de covariance Σ_1.*

3 Approche par projection sur les composantes principales

Dans ce paragraphe, nous adoptons une approche basée sur les projections des τ_ℓ sur des axes principaux, c'est-à-dire des vecteurs propres unitaires de l'opérateur de covariance

V de X, donné par $V = \mathbb{E}\left((X-m)\otimes(X-m)\right)$. Soit $(u_i)_{i\geq 1}$ un système orthonormé de vecteurs propres de V tel que u_i est associé à la $i-ème$ plus grande valeur propre λ_i. On suppose que l'on a pour tout $i \geq 1, \lambda_i > \lambda_{i+1} > 0$. Pour un entier $p \in \mathbb{N}^*$ fixé, on considère $S(p) = \sum_{i=1}^{p}\sum_{\ell=1}^{q} p_\ell \lambda_i^{-1} \langle \tau_\ell, u_i\rangle^2$; il est facile de voir que si $\mathcal{H}_0$ est vraie, on a $S(p) = 0$. Par conséquent, on peut effectuer le test de $\mathcal{H}_0$ contre $\mathcal{H}_1$ en se basant sur un estimateur convergent de $S(p)$. Considérons l'opérateur de covariance empirique $\widehat{V}^n = \frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)} - \overline{X}^n\right)\otimes\left(X^{(i)} - \overline{X}^n\right)$, et $(\widehat{u}_i)_{i\geq 1}$ un système orthonormé tel que $\widehat{u}_i$ soit le vecteur propre de $\widehat{V}^n$ associé à la $i-$ème plus grande valeur propre $\widehat{\lambda}_i$. Pour $p \in \mathbb{N}^*$ suffisamment grand, on définit comme statistique de test la variable aléatoire $\widehat{S}^{(n)}(p) = \sum_{i=1}^{p}\sum_{\ell=1}^{q} p_\ell^n \widehat{\lambda}_i^{-1}\langle \widehat{\tau}_\ell, \widehat{u}_i\rangle^2$ où $\widehat{\tau}_\ell = \overline{X}_\ell^n - \overline{X}^n$. Pour construire le test correspondant, nous allons déterminer la loi asymptotique de $\widehat{S}^{(n)}(p)$ sous $\mathcal{H}_0$. Pour cela, introduisons les opérateurs suivants :

$$\begin{aligned} V^{(\ell)} &= \mathbb{E}\left((X-m_\ell)\otimes(X-m_\ell)\,|Y=\ell\right) \\ \Theta_{j\ell} &= \delta_{j\ell}p_j V^{(j)} + p_j p_\ell\left(V - V^{(j)} - V^{(\ell)}\right) \end{aligned}$$

où $(j,\ell)\in F^2$ et $\delta_{j\ell}$ désigne le symbole de Kronecker. Ceci nous permet de considérer la matrice carrée de dimension pq suivante :

$$\Sigma_2 = \begin{pmatrix} \sigma_{1111} & \sigma_{1112} & \dots & \sigma_{11pq} \\ \sigma_{1211} & \sigma_{1212} & \dots & \sigma_{12pq} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{pq11} & \sigma_{pq12} & \dots & \sigma_{pqpq} \end{pmatrix},$$

où $\sigma_{ijk\ell} = \langle u_i, \Theta_{j\ell}u_k\rangle$. Soient $D_V(p) = diag\left(\lambda_i; 1\leq i\leq p\right)$ et $\Gamma_2 = \left(D_V(p)\right)^{-1}\otimes^K \Delta^{-1}$, où $\otimes^K$désigne le produit de Kronecker. Nous avons alors :

Théorème 2. *Sous $\mathcal{H}_0$, $n\widehat{S}^{(n)}(p)$ converge en loi, lorsque $n\to+\infty$, vers $Q_2 = M_2^T\Gamma_2 M_2$ où M_2 est un vecteur aléatoire gaussien, dans $\mathbb{R}^{pq}$, centré et de matrice des covariances égal à Σ_2.*

Remarquons que Q_1 et Q_2 sont des formes quadratiques d'un vecteur gaussien. On peut par conséquent approcher leur fonction de répartition en appliquant des formules données dans Mathai et Provost (1992) et qui impliquent les opérateurs Γ_1, Σ_1 et Γ_2, Σ_2 respectivement. Ces opérateurs étant inconnus dans la pratique, on les remplace par des estimateurs convergents.

4 Application aux variables fonctionnelles

On se place dans le cas où X est à valeurs dans un espace de Hilbert fonctionnel; il est alors noté $X = (X_t)_{t\in I}$, où I est un intervalle $[a, b]$. On suppose que les $X^{(i)}$ sont observés en des points $t_1, t_2, ..., t_{L_1}$ tels que $a = t_1 < t_2 < ... < t_{L_1} = b$. Sur cette base, on veut appliquer les méthodes introduites dans les paragraphes précédents.

4.1 Application de la première méthode

Lorsque l'on prend H comme l'espace de Sobolev $H^{r-1}(I)$, on considère la projection de X sur le sous-espace engendré par les fonctions B-splines d'ordre r, notés $B_1^{(r)}, ..., B_p^{(r)}$ et calculés sur des noeuds $s_1, ..., s_{L_2}$ vérifiant $a = s_1 < s_2 < ... < s_{L_2} = b$; on a alors nécessairement $p = r + L_2$. Le choix des noeuds est arbitraire mais effectué de telle sorte que le pas de la subdivision induite soit suffisamment petite pour permettre une bonne approximation. Le calcul des $\left\langle \phi_k, X^{(i)} \right\rangle$, où $\phi_k = B_k^{(r)}$, s'effectue par l'utilisation des formules classiques pour l'approximation des dérivées successives et d'une formule de quadrature pour l'approximation de l'intégrale.

Lorsque H est pris comme l'espace $L^2(I)$, nous considérons une analyse multirésolution (MRA) $\{V_n\}_{n\in\mathbb{Z}} \in L^2(I)$. Soit $(j_0, J) \in \mathbb{Z}^2$ tel que $J > j_0$, V_J est le sous-espace de dimension finie engendré par les fonctions d'échelles et d'ondelettes associées à cette MRA. Les fonctions ϕ_k sont pris dans la base de V_J et les produits scalaires $\left\langle \phi_k, X^{(i)} \right\rangle$ sont calculés en utilisant la transformation discrète d'ondelettes.

4.2 Application de la deuxième méthode

Soit $\mathbb{X} = \left(X^{(i)}(t_j)\right)_{1\leq i\leq n, 1\leq j\leq L}$ la matrice de dimension $n \times L$ qui contient toutes les observations de X. On procède à une analyse en composantes principales fonctionnelles de cette matrice en utilisant la bibliothèque *fda* du logiciel *R*. Cette analyse nous fournit les valeurs propres $\widehat{\lambda}_i$ et les composantes principales $\widehat{c}_i(k) = \left\langle X^{(i)}, \widehat{u}_k \right\rangle$. Ceci permet d'appliquer la deuxième méthode proposée.

5 Simulations

Les simulations ont permi d'évaluer la performance des methodes proposées. Le modèle suivant a été utilisé : $I = [0, 1]$, $X(t) = m(t) + e(t)$ où $e(t) \rightsquigarrow N(0, 1)$ et $m(t) = \cos(t + Y\pi\gamma/q)$, q étant le nombre de groupes considérés et $\gamma \in [0, 1]$. Pour différentes tailles d'échantillons $n = 25, 50, 100, 200, 300, 400, 500, 1000$ nous avons effectué 1000 répetitions pour évaluer le niveau et la puissance empiriques, et étudier l'influence du paramètre p. Une étude comparative avec des méthodes existantes (Fan & Lin (1998), Cuevas et *al.* (2004)) a été également été effectuée.

References

[1] Cuevas, A., Febrero, M., Fraiman, R. (2004), An anova test for functional data. *Comput. Statist. Data Anal.* **47**, 111-122.

[2] Dauxois, J., Nkiet, G., M. (1997), Canonical analysis of two Euclidean subspaces and its applications. *Linear Algebra Appl.* **264**, 355-388.

[3] Dauxois, J., Romain, Y., Viguier, S. (1994), Tensor products and statistics. *Linear Algebra Appl.* **210**, 59-88.

[4] Fan, J.,Lin,S.K. (1998), Test of Significance when data are curves. *J.Amer. statist. Assoc.* **93**, 1007-1021.

[5] Mathai, A.M., Provost, S.B. (1992), Quadratic forms in random variables: Theory and applications. Dekker.

Réponse fonctionnelle en cinétique : modélisation et prédiction

François WAHL[a], Matthieu CANAUD[a], Céline HELBERT[b], Laurent CARRARO[c]

[a] *IFP Énergies nouvelles, BP 3 69360 Solaize France*
[b] *Université Joseph Fourier, Tour IRMA, BP 53 38041 Grenoble Cedex 9 France*
[c] *Telecom Saint-Étienne, 23 rue Dr Paul Michelon 42023 Saint Etienne Cedex 2 France*

Résumé Il arrive souvent que le modèle postulé pour représenter un système physique ou chimique ne convienne que partiellement. C'est le cas de l'application présentée ici, modèle cinétique de la dépollution des fumées en sortie d'un moteur diesel. Nous montrons comment cette difficulté a été levée, en autorisant les paramètres du modèle à varier suivant les entrées, en s'inspirant de la méthodologie des 'varying coefficient models'. Le modèle ainsi estimé intègre entièrement la spécificité du modèle cinétique tout en l'adaptant localement aux paramètres d'entrée.

Abstract
It can happen that a model representing a chemical or a physical system works imperfectly and is only partially correct. It is the case of the example in this paper, kinetic model of smoke depollution from the output of diesel engine. We show how this difficulty has been solved, by letting the parameters of the model vary along the entries, following the methodology of 'varying coefficient models'. The new model takes into account all the knowledge of the initial kinetic model, while locally adapting it to the entries.

mot-clefs : Réponse fonctionnelle, Modèle à coefficients variables, Régression non linéaire

keywords : Functional response, Varying Coefficient Model, Non linear Regression

1 Introduction

Il arrive souvent que le modèle postulé pour représenter un système physique ou chimique ne convienne que partiellement. C'est le cas du système présenté ci-dessous, et nous montrons comment cette difficulté a été levée, en autorisant les paramètres du modèle à varier suivant les entrées, en s'inspirant de la méthodologie des 'varying coefficient models' (VCM).

Dans un problème de régression fonctionnelle, les entrées ou les sorties ou les deux peuvent être des fonctions (voir Ferraty et Vieu (2006), Ramsay et Silverman(2005)). Une sortie fonctionnelle peut être décomposée en une tendance centrale additionnée d'un processus alétoire gaussien, comme dans Chiou et al (2003) ou Nerini et al (2010). Cependant, cette démarche ne s'applique pas à notre problème du fait de la non-stationarité de

la réponse (Canaud et al 2009), même en considérant que la moyenne conditionnelle du processus est précisément le modèle initial. Dans ce papier nous nous intéressons d'abord à la démarche de Li et al (2005) dans le cas de sorties linéaires fonctionnelles. Nous étendons cette démarche en faisant varier les coefficients du modèle comme dans le cas de VCM (Fan et al 1999). Enfin, nous l'adaptons en proposant une extension non intrusive du modèle initial.

Dans une première partie, nous introduisons le système expérimental étudié. Après un tour d'horizon des méthodes possibles nous étendons la méthodologie des VCM à notre problème, tout d'abord en nous ramenant à un cas linéaire, puis en adaptant directement le modèle et nous montrons les résultats obtenus.

2 Système expérimental

Nous étudions un système de dépollution catalytique des fumées produites par les moteurs Diesel, pour lequel un modèle a été construit après que les experts aient posé un ensemble de réactions chimiques, et postulé des mécanismes cinétiques.

Le piège à NOx est un système catalytique destiné à dépolluer les gaz d'échappement en sortie des moteurs diesel, c'est à dire à en éliminer les hydrocarbures imbrulés (HC), les monoxydes de carbone (CO) et les oxydes d'azote (NOx). Le piège se présente comme une structure cylindrique poreuse imprégnée de métaux catalyseurs à travers laquelle les gaz passent et réagissent. Nous ne nous intéressons ici qu'aux réactions d'oxydation du CO et des HC.

Les essais expérimentaux sont réalisés sur un appareillage appelé BGS (acronyme de 'Banc Gaz Synthétique') censé représenter la réalité qui comprend, pour simplifier exagérément, deux éléments : un four pour porter les gaz en entrée à une température contrôlée et un réacteur contenant une carotte de catalyseur. En sortie, les gaz sont recueillis et analysés. Ici, nous ne nous intéressons qu'à deux réponses, les teneurs en CO et en HC en sortie. Comme la température T varie continûment au cours d'un essai, ces réponses sont des courbes dépendant de T.

3 Modèle et VCM

Un modèle a été proposé pour rendre compte du comportement expérimental, et se schématise par l'équation :

$$y(T) = f(x, T, \boldsymbol{\beta}), \tag{1}$$

où x est un vecteur à p éléments contenant la composition des gaz en entrée, T est la température de ces gaz, $\boldsymbol{\beta}$ est un vecteur de q paramètres, $y(T)$ est un vecteur à $r = 2$ éléments qui contient les réponses à la température T, f est un système d'équations différentielles ordinaires résolues numériquement.

Pour ajuster ce modèle aux n résultats expérimentaux disponibles et estimer les paramètres du vecteur, la première idée naturelle est de déterminer par moindres carrés un

vecteur de paramètres globaux, valables quelle que soit l'expérience considérée.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \int \|y_i(T) - f(x_i, T, \boldsymbol{\beta})\|^2 \, dT \tag{2}$$

Malheureusement, tous les essais basés sur l'équation (2) se sont révélés infructueux, montrant que le modèle proposé est insuffisant pour représenter toute la réalité expérimentale, même si les grandes tendances sont respectées.

Une deuxième idée pour exploiter au mieux le modèle (1) est de considérer que celui-ci rend compte d'un effet moyen, qui doit être corrigé d'un effet modélisé comme un processus gaussien $Z_\theta(T)$ de moyenne nulle et dont la matrice de variance-covariance est paramétrée par θ, de telle sorte que $y(T) = f(x, T, \boldsymbol{\beta}) + Z_\theta(T)$. Cette deuxième alternative n'est pas un succès, pour deux raisons : la première est que les réalisations observées de Z ne peuvent être considérées comme stationnaires en T, et la seconde que les courbes obtenues doivent respecter impérativement certaines contraintes. En effet, les sorties, une fois normalisées, doivent être décroissantes toujours inférieures à 1 car le piège traite les polluants mais n'en produit pas et positives car elles représentent des concentrations.

Cependant, essai par essai, le modèle (2) donne des résultats satisfaisants, de sorte qu'on peut définir pour chaque expérience un vecteur de paramètres cinétiques.

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}} \int \|y_i(T) - f(x_i, T, \boldsymbol{\beta})\|^2 \, dT, \tag{3}$$

ce qui revient à faire varier les paramètres en fonction des x_i. Cependant, les $\hat{\boldsymbol{\beta}}_i$ obtenus sont dispersés, deux expériences proches pouvant conduire à des $\hat{\boldsymbol{\beta}}_i$ très différents.

Les approches alternatives que nous avons examinées dans un premier temps - Li et al (2005) et Fan et Zhang (1999) - quoique d'origine très différentes, proposent réponse par réponse le modèle local linéaire suivant :

$$y = x^t \beta(T) + \epsilon, \tag{4}$$

avec $\mathbb{E}(\epsilon) = 0$ et $\mathbb{E}(\epsilon^2) = \sigma^2$ (voir Hastie et Tibshirani 1993), c'est à dire que T prend le rôle d'une variable longitudinale. L'information apportée par f n'est plus prise en compte. Le vecteur de paramètres $\boldsymbol{\beta}$ est donc dans ce cas de même dimension que x et a donc p composantes qui n'ont plus de significations cinétiques particulières.

Li, Sudjanto et Zhang (2005) proposent un algorithme simple à mettre en œuvre, qui consiste à obtenir une première estimation par moindres carrés des $\boldsymbol{\beta}(T)$ pour chaque température indépendamment, puis à lisser les estimations obtenues dans une seconde étape, par exemple par P-splines (Eilers et Marx 1996).

Fan et Zhang (1999) suggèrent quant à eux une méthode basée sur les polynômes locaux. Leur formalisme peut être adapté aux cas de sorties fonctionnelles. Dans cette approche, les $\beta(T)$ se représentent localement comme des polynômes de degré n_d, ce

degré étant généralement choisi inférieur à 3. En un point t voisin de T, on écrit : $\boldsymbol{\beta}(t) = \sum_{d=1}^{n_d} b_d(t-T)^d$.

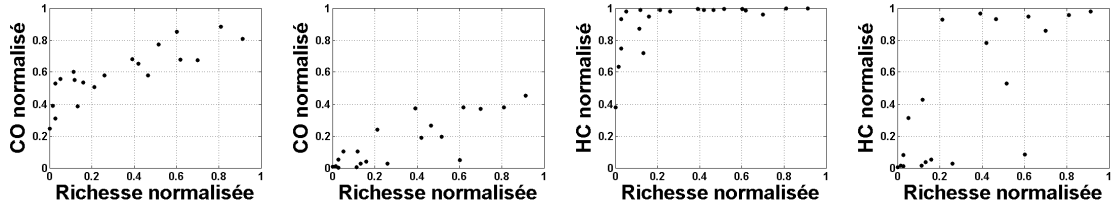
La détermination des coefficients b_d du polynôme se fait réponse par réponse et le critère à minimiser devient pour chaque réponse (l'indice r correspondant à celle-ci est omis pour simplifier) :

$$\hat{\boldsymbol{\beta}}(T) = \arg\min_{b} \sum_{i=1}^{n} \int (y_i(t) - x_i^t \boldsymbol{\beta}(t))^2 K_h(t-T)dT, \tag{5}$$

où K_h est une fonction noyau, dépendant d'un hyper-paramètre h. Usuellement, on choisit pour K le noyau d'Epanechnikov d'expression $K_h(u) = \frac{3}{4}(1-\frac{u^2}{h^2})_+$ pour u variant sur [-1,1].

Quand on estime $\hat{\boldsymbol{\beta}}(T)$, l'introduction du noyau permet de pondérer l'importance des observations puisque plus les points sont éloignés de T, moins ils ont d'influence, les points distants de plus de h n'intervenant plus dans l'estimation. De plus, un choix correct de h permet de lisser l'estimation.

L'examen des points expérimentaux à T constant en fonction de x permet de valider l'idée d'une dépendance longitudinale en température pour la concentration en CO (voir figures 1 (a) et (b)) mais pas pour la concentration en HC (figures 1 (c) et (d)).



(a) Sortie CO en $T = 0.4$(b) Sortie CO en $T = 0.5$(c) Sortie HC en $T = 0.4$(d) Sortie HC en $T = 0.5$

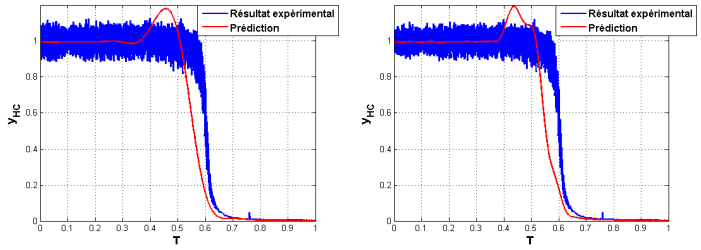
Figure 1 – Nuages de points des réponses normalisées CO et HC selon la richesse pour deux températures

Comme on pouvait s'y attendre, les résultats obtenus, que ce soit en suivant la méthode de Li ou de Fang, sont acceptables pour la première réponse mais ne sont pas satisfaisants pour la seconde (voir figure 2), les courbes pouvant présenter au moment du démarrage effectif de la réaction un sursaut choquant pour l'œil d'un expert.

4 Extension au cas non linéaire

Les méthodes précédentes ont le grave inconvénient de ne pas tenir compte de la connaissance des experts en ignorant le modèle (1), et de forcer un caractère linéaire quelle que soit la température considérée (voir figure 1d).

Pour remédier à ces défauts, nous proposons deux méthodes d'estimation des coefficients, en s'inspirant des méthodes précédentes. La première consiste à résoudre un



(a) Prédiction HC par la méthode de Li *et al*

(b) Prédiction HC par la méthode de Fang et Zhang

Figure 2 – Exemples de résultats pour les deux méthodes

moindre carré pondéré et à poser :

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \int \|y_i(T) - f(x_i, T, \boldsymbol{\beta})\|^2 K_h(x_i, x) dT \tag{6}$$

Ici, la fonction noyau dépend des x au lieu de T. Dans le cas d'un noyau gaussien, le paramètres h est un vecteur de même dimension p que x.

La seconde méthode, dans la lignée de Hastie et Tibshirani (1993), propose de remplacer les $\hat{\boldsymbol{\beta}}_i$ du modèle (2) par des coefficients dont les estimations sont lissées. Pour ce faire, Hastie et Tibshirani utilisent les B-splines de lissage. Nous avons choisi d'employer la méthode 'Kernel Ridge Regression' (KRR), détaillée par exemple dans Saunders *et al* (1998).

Si pour chaque expérience i, le vecteur b_i des "vrais" paramètres non lissés du modèle était connu, les coefficients estimés et lissés $\hat{\boldsymbol{\beta}}_i$ se prédiraient en KRR par la relation linéaire en b_i : $\hat{\boldsymbol{\beta}}_i = k^t(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} b_i$, où $K_h(u, v)$ est une fonction noyau, k désigne le vecteur de dimension n dont chaque composante vaut $K(x_i, x)$ et $\boldsymbol{K}$ est la matrice de composantes $K(x_i, x_j)$, $\boldsymbol{I}$ est l'identité d'ordre n, λ est un hyper paramètre.

Cependant, les b_i sont inconnus et nous les collectons dans un vecteur $\hat{b}$ qui contient donc $n * q$ composantes pour n expériences et q paramètres cinétiques par expérience. Le critère à estimer ressemble au critère (2), sauf que la minimisation porte maintenant sur b, à partir duquel on reconstruit chaque $\boldsymbol{\beta}_i$:

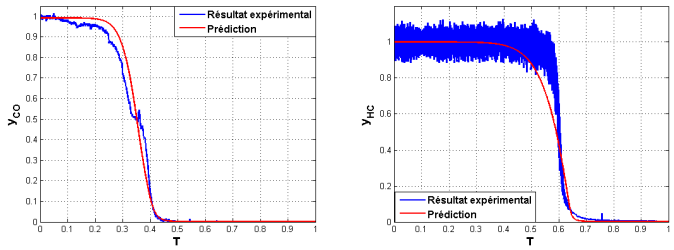
$$\hat{b} = \arg\min_{b} \sum_{i=1}^{n} \int \|y_i(T) - f(x_i, T, \boldsymbol{\beta}_i)\|^2 dT \tag{7}$$

Dans le cas de notre application, $q = 5$ et $n = 20$ et la minimisation (7) porte donc sur $n * q = 100$ variables au lieu des 5 initiales.

Les résultats obtenus par ces deux méthodes (6) et (7) sont comparables et illustrés sur la figure 3. La seconde méthode offre l'avantage de déterminer l'équation d'une surface de

réponse pour représenter l'évolution des paramètres, qu'il suffit d'évaluer pour déterminer un nouveau point, alors que dans (6) une optimisation est nécessaire chaque fois qu'on veut déterminer $\hat{\boldsymbol{\beta}}$ pour un nouveau x. Ces deux méthodes ne nécessitent que d'interfacer le calcul du modèle (1), et sont donc non intrusives.

Les démarches proposées (6) et (7) emploient des méthodes analogues à celles des VCM, et en étendent l'application à un cadre non linéaire.



(a) Prédiction *CO* (b) Prédiction *HC*

Figure 3 – Prédictions des sorties normalisées par le modèle (7)

Bibliographie

[1] Ferraty F. et Vieu P. (2006) *Nonparametric functional data analysis : methods, theory, applications and implementations*, Springer-Verlag, London.

[2] Ramsay J.O. et Silverman B.W. (2005) *Functional data analysis*, 2nd edition, Springer, New York.

[3] Chiou J.M. et Müller H.G. et Wang, J.L. (2003) *Functional quasi-likelihood regression models with smooth random effects*, Journal of the Royal Statistical Society : Series B (Statistical Methodology), 65 :405-423.

[4] Nerini D., Monestiez P. et Mantéa C. (1997) *Cokriging for spatial functional data*, Journal of Multivariate Analysis, 101 :409-418.

[5] Canaud M., Wahl F., Helbert, C. et Carraro L. (2009) *Design of experiments for smoke depollution from the output of diesel engine*, Congrès ENBIS-EMSE, Saint-Étienne.

[6] Li R., Sudjianto A. et Zhang Z. (2005) *Modeling computer experiments with functional response*, SAE SP, 1956 :309-316.

[7] Fan J. et Zhang, W. (1999) *Statistical estimation in varying-coefficient models*, The Annals of Statistics, 27(5) :1491-1518.

[8] Hastie T. et Tibshirani R. (1993) *Varying-Coefficient Models*, Journal of the Royal Statistical Society. Series B (Methodological), 55(4) :757-796.

[9] Eilers P.H.C. et Marx B.D. (1996) *Flexible smoothing with B-splines and penalties*, Statist. Sci., 25(1) :89-121.

[10] Saunders C., Gammerman A. et Vovk V. (1998) *Ridge Regression Learning Algorithm in Dual Variables*, Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin.

Régression ridge à noyau pour des variables explicatives et d'intérêts fonctionnelles

Hachem Kadri[1], Philippe Preux[1,2] & Emmanuel Duflos[1,3]

[1]*Equipe-projet SequeL, INRIA Lille - Nord Europe, Villeneuve d'Ascq, France*
[2]*LIFL/CNRS, Université de Lille, Villeneuve d'Ascq, France*
[3]*LAGIS/CNRS, Ecole Centrale de Lille, Villeneuve d'Ascq, France*

Résumé

Dans cet article, on s'intéresse à la régression fonctionnelle dans le cas où les variables explicatives ainsi que les variables d'intérêts sont de dimension infinie et représentées par des fonctions. Nous introduisons une méthode d'estimation non-paramétrique de la fonction de régression, extension de la régression ridge à noyau au domaine de l'analyse des données fonctionnelles, basée sur la généralisation de la théorie des espaces de Hilbert à noyaux reproduisants (EHNR). L'originalité du travail réside essentiellement dans le choix du noyau reproduisant construit à partir de l'opérateur intégral et la proposition d'une procédure pour l'inversion de la matrice noyau à blocs opérateurs.

Abstract

This paper deals with the problem of functional regression where covariates as well as responses are functions. Basic concepts of reproducing kernel Hilbert space (RKHS) theory are extended to the domain of functional data analysis and a functional kernel ridge regression algorithm is provided. Our main results demonstrate how to build the reproducing kernel from the integral operator and how to invert the corresponding block-operator kernel matrix.

Mots clés. Régression ridge, analyse des données fonctionnelles, méthodes à noyaux, variable d'intérêt fonctionnelle, espace de Hilbert à noyaux reproduisants, noyau à valeurs opérateurs.

1. Introduction

La régression est une problématique de recherche qui met en évidence le besoin de développer des méthodes statistiques et d'apprentissage automatique afin d'apporter une meilleure compréhension des processus physiques, biologiques et naturels. Elle a pour objectif de décrire les relations possibles entre des variables observées et mesurées et permet de produire un modèle permettant de prédire les valeurs prises par une variable qu'on désire expliquer à partir d'une série de variables explicatives continues et/ou catégorielles. Ces dernières années plusieurs travaux de recherche, sur le plan théorique et algorithmique, ont été menés pour développer des méthodes de régression et de régularisation. Plusieurs méthodes de régularisation par des fonctions convexes

se sont avérées très efficaces. On peut citer par exemple : le Lasso (Tibshirani, 1996), l'Elastic Net (Zou et Hastie, 2005) et les SVR (Support Vector Regression) (Drucker et al., 1997). Ces méthodes ont été développées pour traiter des données discrètes, alors qu'actuellement, dans de nombreux domaines, les quantités mesurées ne sont plus des éléments de $\mathbb{R}^d$ mais des objets plus complexes: courbes, images, etc. Dans ce sens, on s'intéresse dans cet article à la régression sur des données fonctionnelles (Ramsay et Silverman, 2005), plus précisemment le cas où les variables explicatives et à expliquer sont de dimension infinie et donc représentées par des fonctions. Les développements récents, dans le domaine de l'analyse de données fonctionnelles, des méthodes de régression non-paramétriques pour variables fonctionnelles offrent de nombreuses perspectives. Le spectre d'applications de ces méthodes est largement étendu, on peut citer par exemple l'analyse et la quantification du risque d'un investissement en finance, l'analyse des données d'expressions de gènes en biologie, l'étude de l'interaction entre les variables climatiques en science environnementale et la localisation des activations cérébrales au cours d'une tâche comportementale à partir des données d'IRM fonctionnelle (Sood et al., 2009; Escabias et al., 2005; Ramsay et Silverman, 2002).

La statistique fonctionnelle a connu un très important développement ces dernières années, donnant naissance à plusieurs méthodes paramétriques (Ramsay et Silverman, 2005) et d'autres non-paramétriques (Ferraty et Vieu, 2006) permettant l'analyse des données fonctionnelles. Les récents articles de Preda (2007), Lian (2007) et Kadri et al. (2010b) mettent en évidence l'apport de la théorie des espaces de Hilbert à noyaux reproduisants (EHNR) pour développer des méthodes d'estimation non-paramétriques appropriées aux données fonctionnelles. Dans ce contexte, les propriétés et caractéristiques des EHNR vérifiées dans le cas scalaire sont étendues pour inclure le cas où les variables explicatives et les variables d'intérêts sont des fonctions. Le théorème de représentant est généralisé pour montrer que l'opérateur de régression solution d'un problème de minimization régularisé s'ecrit sous forme d'une combiniason de fonctions noyaux à valeurs opérateurs. Le présent papier vise à proposer une extension de la régression ridge à noyau au cas fonctionel en se basant sur les EHNR à valeurs fonctions. Un intért particulier est porté à la construction du noyau reproduisant à partir du l'opérateur intégral et à l'inversion de la matrice noyau à blocs opérateurs.

2. EHNR à valeurs fonctions

Un modèle non-paramétrique de régression fonctionnelle avec variables d'intérêts fonctionnelles y_i s'écrit sous la forme suivante:

$$y_i(t) = f(x_i(s)) + \epsilon_i(t), \quad s \in I_s,\ t \in I_t, \quad i = 1, \ldots, n, \tag{1}$$

avec x_i les variables explicatives, ϵ_i une suite de variable aléatoire et f la fonction de régression à estimer. La méthode de régression présentée dans ce travail est basée sur l'approximation de l'opérateur de régression f dans un EHNR $\mathcal{F}$ construit à partir d'un

noyau à valeurs opérateurs positive K. Cette section introduit les EHNR à valeurs fonctions utilisés pour approximer des opérateurs linéaires (Kadri et al., 2010b).

Soient $\mathcal{G}_x$ et $\mathcal{G}_y$ des espaces de Hilbert de dimensions infinies et $\mathcal{F}$ l'espace des opérateurs linéaires de $\mathcal{G}_x$ à valeurs dans $\mathcal{G}_y$, muni d'un produit scalaire $\langle .,. \rangle_{\mathcal{F}}$. $\mathcal{L}(\mathcal{G}_y)$ est l'ensemble des opérateurs linéaires bornés de $\mathcal{G}_y$ dans $\mathcal{G}_y$.

Définition 1 *(EHNR à valeurs fonctions)*
Un espace de Hilbert $\mathcal{F}$ de fonctions de $\mathcal{G}_x$ dans $\mathcal{G}_y$ est un espace de Hilbert à noyaux reproduisants s'il existe un noyau à valeurs opérateurs $K_{\mathcal{F}}(w,z) : \mathcal{G}_x \times \mathcal{G}_x \longrightarrow \mathcal{L}(\mathcal{G}_y)$ telque:

i. la fonction $z \longmapsto K_{\mathcal{F}}(w,z)g$ est dans $\mathcal{F}$, $\forall z \in \mathcal{G}_x$, $w \in \mathcal{G}_x$, $g \in \mathcal{G}_y$,

ii. $\forall f \in \mathcal{F}, \langle f, K_{\mathcal{F}}(w,.)g \rangle_{\mathcal{F}} = \langle f(w), g \rangle_{\mathcal{G}_y}$ (propriété reproduisante).

Définition 2 *(noyau à valeurs opérateurs)*
Un noyau $K_{\mathcal{F}}(w,z)$ à valeurs opérateurs est une fonction de $\mathcal{G}_x \times \mathcal{G}_x$ dans $\mathcal{L}(\mathcal{G}_y)$

- *$K_{\mathcal{F}}$ est Hermitien si $K_{\mathcal{F}}(w,z) = K_{\mathcal{F}}(z,w)^*$, avec $K_{\mathcal{F}}(z,w)^*$ est l'opérateur adjoint de $K_{\mathcal{F}}(z,w)$*
- *$K_{\mathcal{F}}$ est positif s'il est Hermitien et $\forall r \in \mathbb{N}^+$ et $\forall \{(w_i, u_i)_{i=1,\ldots,r}\} \in \mathcal{G}_x \times \mathcal{G}_y$, $\sum_{i,j} \langle K_{\mathcal{F}}(w_i, w_j)u_i, u_j \rangle_{\mathcal{G}_y} \geq 0$.*

Théorème 1 *(bijection entre les EHNR à valeurs fonctions et les noyaux à valeurs opérateurs)*
Un noyau $K_{\mathcal{F}}(w,z)$ de $\mathcal{G}_x$ dans $\mathcal{L}(\mathcal{G}_y)$ est le noyau reproduisant d'un espace de Hilbert $\mathcal{F}$, si et seulement si $K_{\mathcal{F}}(w,z)$ est positif.

Pour la démonstration de ce théorème ainsi que la construction d'un EHNR autour d'un noyau à valeurs opérateurs et l'unicité de cet espace et du noyau crées, nous renvoyons le lecteur au rapport de recherche de Kadri et al. (2010a).

3. Régression ridge à noyau fonctionnelle

Dans cette section, nous étudions l'estimation de l'opérateur de régression f, défini par l'équation (1), dans un EHNR $\mathcal{F}$ construit autour d'un noyau à valeurs opérateurs K. La régression ridge consiste à résoudre un problème de minimisation avec une régularisation au sens de Tikhonov qui combine une fonction coût L^2 et une régularisation L^2. Une estimation de f dans l'espace de Hilbert à noyaux reproduisants $\mathcal{F}$ est l'opérateur f^* solution du problème de minimisation suivant:

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \|y_i - f(x_i)\|^2_{\mathcal{G}_y} + \lambda \|f\|^2_{\mathcal{F}} \tag{2}$$

Utilisant le théorème de représentant dans le cas fonctionnel (Kadri et al., 2010a), la solution de ce problème est donnée par la formule suivante:

$$f^*(x) = \sum_{j=1}^{n} K(x, x_j)\beta_j, \quad \beta_j \in \mathcal{G}_y \tag{3}$$

Le problème (2) revient donc à résoudre un problème de minimisation sur $\beta_v \in (\mathcal{G}_y)^n$ le vecteur contenant les fonctions $(\beta_i)_{i=1,\ldots,n}$ au lieu de l'opérateur f (voir ésquation (4)).

$$\min_{\beta_v \in (\mathcal{G}_y)^n} \sum_{i=1}^{n} \|\beta_i - \sum_{j=1}^{n} K(x_i, x_j)\beta_j\|^2_{\mathcal{G}_y} + \lambda \sum_{i,j}^{n} \langle K(x_i, x_j)\beta_i, \beta_j \rangle_{\mathcal{G}_y} \tag{4}$$

Contrairement aux travaux de Lian (2007) et Kadri et al. (2010b) qui résolvent le problème (4) en discrétisant les fonctions x_i, y_i et β_i, nous présentons dans cet article une solution analytique fonctionnelle de ce problème. Utilisant la dérivée directionnelle, nous obtenons que β_v solution du problème satisfait le système d'équations d'opérateurs linéaires suivant:

$$(\mathcal{K} + \lambda I)\beta_v = y_v \tag{5}$$

avec $\mathcal{K} = K(x_i, x_j)_{i,j=1}^{n}$ est une matrice par blocs opérateurs ($\mathcal{K}_{ij} \in \mathcal{L}(\mathcal{G}_y)$) et $y_v \in (\mathcal{G}_y)^n$ est le vecteur de fonctions $(y_i)_{i=1}^{n}$. L'équation (5) est l'extension de la solution de la régression ridge à noyau au cas fonctionnel. La principale difficulté rencontrée dans cette extension se rapporte à l'inversion de la matrice noyau. En effet, $\mathcal{K}$ est une matrice d'opérateurs dont l'inverse n'est pas toujours possible à calculer. Pour surmonter cette difficulté, nous optons pour une décomposition en valeurs propres de la matrice $\mathcal{K}$ construite à partir des noyaux à valeurs opérateurs ayant la forme suivante:

$$K(x_i, x_j) = G(x_i, x_j)T, \quad \forall x_i, x_j \in \mathcal{G}_x \tag{6}$$

avec G une fonction à valeurs réelles et T est un opérateur dans $\mathcal{L}(\mathcal{G}_y)$. Ce choix est inspiré de la méthode de construction de noyaux à valeurs matricielles proposée par Micchelli et Pontil (2005). Le choix de l'opérateur T dépend du contexte. Lian (2007) suggère d'utiliser l'opérateur identité, alors que Kadri et al. (2010b) proposent un noyau fonctionnel construit à partir de l'opérateur multiplication. Dans ce dernier cas, la méthode proposée peut être vue comme étant une extension au cas non linéaire des méthodes de régression basées sur le modèle linéaire fonctionnel concurrent (Ramsay et Silverman, 2005). Dans cet article, on s'intéresse à des noyaux reproduisants construits à partir de l'opérateur intégral puisque le modèle linéaire de régression fonctionnelle à variable d'intérêt fonctionnelle est basé sur cette opérateur (Crambes et Mas, 2010; Ramsay et Silverman, 2005). Nous considérons le noyau à valeurs opérateurs suivant:

$$(K(x_i, x_j)y)(t) = G(x_i, x_j) \int_{\Omega} e^{-|t-s|} y(s)ds, \quad y \in \mathcal{G}_y, \quad \{s, t\} \in \Omega \tag{7}$$

Notons qu'un noyau similaire a été proposé par Caponnetto et al. (2008) pour des espaces de fonctions de $\mathbb{R}$ dans $L^2(\mathbb{R})$.

La matrice noyau associée à des noyaux qui vérifient la condition (6) peut s'écrire sous forme d'un produit de Kronecker entre une matrice $\mathcal{G} = G(x_i, x_j)_{i,j=1}^n$ dans $\mathbb{R}^{n\times n}$ et un opérateur $T \in \mathcal{L}(\mathcal{G}_y)$

$$\mathcal{K} = \begin{pmatrix} G(x_1,x_1)T & \dots & G(x_1,x_n)T \\ \vdots & \ddots & \vdots \\ G(x_n,x_1)T & \dots & G(x_n,x_n)T \end{pmatrix} = \mathcal{G} \otimes T \tag{8}$$

Dans ce cas, la décomposition en valeurs propres de $\mathcal{K}$ est déterminée à partir des décompositions spectrales de la matrice $\mathcal{G}$ et de l'opérateur T. Soient θ_i et z_i les valeurs propres et les vecteurs de fonctions propres de $\mathcal{K}$. L'opérateur inverse $\mathcal{K}^{-1}$ est défini par

$$\mathcal{K}^{-1} e_v = \sum_i \theta_i^{-1} \langle e_v, z_i \rangle z_i, \quad \forall e_v \in (\mathcal{G}_y)^n \tag{9}$$

et les fonctions β_i sont calculées en résolvant le système d'équations (5).

4. Conclusion

Dans ce papier, nous avons présenté une extension de la régression ridge à noyau au cas fonctionnel. Cette extension est basée sur l'estimation non-paramétrique de l'opérateur de régression dans un espace de Hilbert à noyaux reproduisants construits à partir de noyaux à valeurs opérateurs positifs. Nous avons proposé un noyau basé sur l'opérateur intégral et une procédure pour inverser la matrice noyau correspondante permettant de résoudre le problème de minimisation associé à la régression ridge fonctionnelle sans discrétiser les variables explicatives et d'intérêts fonctionnelles.

Remerciements

H. Kadri est soutenu par le contrat jeune chercheur N° 4297 de la région Nord-Pas de Calais.

Bibliographie

[1] Caponnetto, A., Micchelli, C. A., Pontil, M., et Ying, Y. (2008) Universal multi-task kernels. *Journal of Machine Learning Research*, 68,1615–1646.

[2] Crambes, C. et Mas, A. (2010) Prédiction en régression linéaire fonctionnelle avec variable d'intérêt fonctionnelle. *42èmes Journées de Statistique*.

[3] Drucker, H., Burges, C., Kaufman, L., Smola, A., et Vapnik, V. (1997) Support vector regression machines. *Advances in Neural Information Processing Systems 9*, 155–161.

[4] Escabias, M., Aguilera, A. et Valderrama, M. (2005) Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 95–107.

[5] Ferraty, F. et Vieu, P. (2006) *Nonparametric functional data analysis*, New York: Springer.

[6] Kadri, H., Duflos, D., Davy M., Preux, P. et Canu, S. (2010a) A general framework for nonlinear functional regression with reproducing kernel Hilbert spaces. *Rapport de Recherche INRIA*, RR-6908.

[7] Kadri, H., Preux, P., Duflos, D., Canu, S. et Davy M. (2010b) Nonlinear functional regression: a functional RKHS approach. *in Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics (AI & Stats)*, JMLR: W&CP 9, 374–380.

[8] Lian, H. (2007) Nonlinear functional models for functional responses in reproducing kernel hilbert spaces. *Canadian Journal of Statistics*, 35, 597–606.

[9] Micchelli, C. A. et Pontil, M. (2005) On learning vector-valued functions. *Neural Computation*, 17, 177–204.

[10] Preda, C. (2007) Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137, 829-840.

[10] Ramsay, J. et Silverman, B. (2002) *Applied functional data analysis*, New York: Springer.

[11] Ramsay, J. et Silverman, B. (2005) *Functional data analysis*, New York: Springer.

[12] Sood, A., James, G. et Tellis, G. (2009) Functional regression: a new model and approach for predicting market penetration of new products. *Marketing Science*, 28(1), 36–51.

[13] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B.*, 58(1), 267–288.

[14] Zou, H. et Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B.*, 67(2), 301–320.

Fiablilité - Incertitudes

Encadrement et estimation parcimonieuse de probabilités de dépassement en sortie d'un code monotone, *Nicolas Bousquet*

We consider the exceedance of a threshold reliability value by the unidimensional output of a computer code G with multivariate probabilistic input U. When G is assumed time-consuming and monotonous with respect to U, the Monotonous Reliability Method, proposed by de Rocquigny (2009) in an engineering context, can provide bounds and crude estimates of the probability of occurence p of such an undesirable event. In the talk associated to this article, the formalization and a technical deepening of this idea are considered, as a large basis for future theoretical and applied studies. Three kinds of results are especially emphasized. First, the bounds themselves remain too crude and conservative estimators of p for a dimension of U upper than 2. Second, a maximum-likelihood estimator of p can be easily built, presenting a high variance reduction with respect to a standard Monte Carlo case, but suffering from conservative bias. Third, the theoretical properties of a family of unbiased estimators of p, based on sequential nested importance samplings, are analyzed. Their supplementary potential improvement requires further studies whose main lines will be discussed. We show that both approaches lead to promising parsimonious estimation algorithms provided a sequential emulation of the limit state (failure) surface, seen as a supervised classification problem, can be made under monotony constraints.

Hiérarchisation des sources d'incertitude vis à vis d'une probabilité de dépassement de seuil - Une méthode basée sur la pondération des lois, *Paul Lemaître and Aurélie Arnaud*

Lors de la réalisation d'une analyse de sensibilité d'un modèle numérique, la plupart des méthodes existantes sont pensées pour le cas où la quantité d'intérêt est une variance. Les quelques méthodes disponibles quand la quantité d'intérêt est une probabilité sont soit basées sur des hypothèses restrictives, soit nécessitent un grand nombre d'évaluations. Cet article présente une méthode basée sur le tirage d'importance pour estimer l'impact de la modification d'un paramètre d'une loi d'entrée sur la probabilité de dépassement d'un seuil par la sortie d'un modèle. Cet outil s'avère être un complément intéressant pour une analyse de sensibilité quand le code est coûteux en temps de calcul, car il ne se base que sur des simulations déjà effectuées.

An Application of the AFT Model based on the Generalized Weibull Distribution for Reliability Analysis of Redundant Systems, *Noureddine Saaidia, Mikhaïl Nikulin and Ramzan Tahir*

In Reliability, the distributions that have unimodal hazard function are not too much, they include : Log-normal, log-logistic, power generalized Weibull and inverse Gaussian distributions. In this study, we give the application of the Generalized Weibull distribution in the analysis

of redundant system with one main unit and one stand-by units. Asymptotic properties of the estimators and asymptotic confidence intervals are obtained.

Réflexions sur l'analyse d'incertitude dans un contexte industriel : information disponible et enjeux décisionnels, *Merlin Keller, Alberto Pasanisi, Khouloud Ghorbel and Eric Parent*

Uncertainty analysis aims at quantifying the uncertainty affecting the value of a quantity of interest, characteristic of a the functioning of a physical system, and related to important decisional issues. Most approaches rely on statistical inference, and are divided into three classes : so-called "plug-in" approaches, providing a point estimate of the quantity of interest ; Bayesian procedures that infer an optimal value of the quantity of interest from a cost function formalizing the decisional stakes, and finally purely descriptive approaches, that describe the uncertainty affecting the quantity of interest. The relevance of each approach varies depending on the goals of the uncertainty analysis, and the amount of data available to the analyst. We illustrate this on a real dataset of a water height and discharge measures.

ENCADREMENT ET ESTIMATION PARCIMONIEUSE DE PROBABILITÉS DE DÉPASSEMENT EN SORTIE D'UN CODE MONOTONE

Nicolas Bousquet

EDF R&D, 6 quai Watier, 78401 Chatou.

Summary. We consider the exceedance of a threshold reliability value by the unidimensional output of a computer code G with multivariate probabilistic input $\boldsymbol{\nu}$. When G is assumed time-consuming and monotonous with respect to $\boldsymbol{\nu}$, a method proposed by de Rocquigny (2009) can provide bounds of the probability of occurence p_f of such an undesirable event. In this paper, the formalization and a technical deepening of this idea are explored. Especially, two class of statistical estimators of p_f are obtained based on sequential nested importance samplings. They can show good properties of robustness and parsimony, ie. variance reduction with respect to a standard Monte Carlo estimator. Promising algorithms can be based on a sequential emulation of the limit state (failure) surface, seen as a supervised classification problem under a monotony constraint, which remains an open problem.

Keywords: structural reliability ; exceedance probability ; computer code ; input uncertainty ; Monte Carlo acceleration ; probability bounds.

Résumé. Soit une variable décrite comme la sortie unidimensionnelle d'un code de calcul déterministe G d'entrées aléatoires, et soit p_f la probabilité que cette variable soit au-dessus d'un certain seuil, évènement supposé indésirable. Quand G est supposé monotone par rapport à ses entrées, une méthode proposée par de Rocquigny (2009) peut fournir des bornes encadrant p_f en fonction d'un plan d'expérience séquentiel. Un approfondissement technique de cette méthode constitue le propos de cet article. Deux classes d'estimateurs statistiques de p_f peuvent être produites en complément des bornes. Ils présentent de bonnes propriétés de parcimonie, c'est-à-dire qu'ils nécessitent pour leur calcul un nombre d'appels au code plus faible qu'un estimateur de Monte Carlo classique. Des algorithmes d'estimation prometteurs peuvent être construits à partir d'une émulation séquentielle de la surface d'état limite associée au dépassement. Il faut pour cela résoudre une suite de problèmes de classification supervisée sous contraintes.

Mots-clés: fiabilité structurelle ; probabilité d'excès ; code de calcul déterministe ; entrées de code incertaines ; Monte Carlo accélérée ; bornes de probabilité.

1 - Contexte

Soit $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ un vecteur de d variables aléatoires uniformes sur $[0,1]$ et indépendantes. Soit G un code de calcul déterministe, continu vis-à-vis de ses entrées et lourd d'exécution, tel que $Z = G(\boldsymbol{\nu})$ soit une variable aléatoire univariée. Soit $Y_{\boldsymbol{\nu}}$ la variable binaire $\mathbb{1}_{\{G(\boldsymbol{\nu})\leq 0\}} = \mathbb{1}_{\{Z\leq 0\}}$. Dans un calcul de propagation d'incertitude, à une reparamétrisation près, le vecteur

$\boldsymbol{\nu}$ peut représenter des variables environnementales (ex: la géographie d'un terrain, les dimensions et le débit d'un cours d'eau) et Z une variable de décision (ex: -Z = hauteur relative de la rivière par rapport au niveau 0 normatif, si G est un code hydraulique aux éléments finis). L'évènement $Y_{\boldsymbol{\nu}} = 1$ est généralement défini comme indésirable (ex: occurence d'une crue), et on cherche donc à estimer de façon parcimonieuse (en minimisant le nombre d'appels au code) sa probabilité d'occurence $p_f = P(G(\boldsymbol{\nu}) \leq 0)$. Il s'agit d'un problème de *fiabilité structurelle*.

On définit l'ordre partiel $\mathbf{x} \succeq \mathbf{y}$ par $x_i \geq y_i$ pour $i = 1, \ldots, d$. Si on suppose que G est *monotone croissant* par rapport à ses entrées selon cet ordre, tout point $\mathbf{x}$ de l'espace $\mathbb{U} = [0,1]^d$ qui est situé en-dessous d'un point $\mathbf{u}$ tel que $y_{\mathbf{u}} = 1$ (toujours selon cet ordre) est nécessairement tel que $y_{\mathbf{x}} = 1$. De même, tout $\mathbf{y}$ de $\mathbb{U}$ situé au-dessus d'un point $\mathbf{u}$ tel que $y_{\mathbf{u}} = 0$ est nécessairement tel que $y_{\mathbf{y}} = 0$. La monotonie de G est par exemple avérée pour certains codes hydrauliques, ou la modélisation de la résistance de structures dans le domaine de la construction de bâtiments (de Rocquigny 2009, Limbourg et al 2010).

On suppose donc pouvoir définir deux ensembles $\mathbb{U}_n^-$ et $\mathbb{U}_n^+$ de points (c'est-à-dire un plan d'expérience) "dominés" par les n sommets de deux unions d'hypercubes situées de part et d'autre de la *surface de défaillance* $\{\boldsymbol{\nu} \in [0,1]^d,\ G(\boldsymbol{\nu}) = 0\}$ (cf. Figure 1). Il faut noter que celle-ci est une fonction décroissante des coordonnées, et que le volume situé sous cette hypersurface est égal à p_f. On note que $\mathrm{Vol}(\mathbb{U}_n^-)$ et $1 - \mathrm{Vol}(\mathbb{U}_n^+)$ constituent donc des bornes inférieure et supérieure pour p_f (de Rocquigny 2009). Celles-ci peuvent être calculées sans difficulté computationnelle pour une dimension d quelconque (Bousquet 2010), de façon exacte (pour d faible), ou approchée avec une précision arbitraire (quelque soit d).

La façon optimale de construire le plan d'expérience correspondant aux sommets des unions hypercubiques est de procéder pas par pas, en plaçant $\boldsymbol{\nu}_k$ dans l'espace non-dominé $\mathbb{U}_{k-1} = \mathbb{U}/(\mathbb{U}_{k-1}^- \oplus \mathbb{U}_{k-1}^+)$ à l'itération k. Ce plan d'expérience peut d'abord être construit de façon déterministe, afin d'agrandir au maximum le volume de l'espace dominé ôté de l'étude. On procède ainsi (Figure 2) jusqu'à obtenir des bornes $(p_{k_0}^-, p_{k_0}^+)$ non triviales (autres que 0 et 1), à une étape $k_0 \geq \log(1/p_f)/d \log 2$.

2 - Limitations et objectifs

Cependant un plan d'expérience déterministe, régulièrement réparti, mène à un coût exploratoire exponentiel en termes de nombre d'essais. Par ailleurs, lorsque d augmente les bornes progressives p_n^- et p_n^+ risquent de rester éloignées et trop conservatives. Plus qu'à utiliser la borne supérieure comme estimateur conservatif de p_f, on cherche plutôt à utiliser ces bornes pour produire une réelle estimation statistique de p_f, parcimonieuse en nombre d'appels à G. Celle-ci émane du choix d'un plan d'expérience *stochastique*, où l'on considère que $\boldsymbol{\nu}_k \sim f_{k-1}(\boldsymbol{\nu})$ où f_{k-1} est une densité de support $\mathbb{U}_{k-1}$. Est dit ici *parcimonieux* un estimateur de variance plus faible qu'un estimateur de Monte Carlo classique $\hat{p_f}_n$, sans biais et

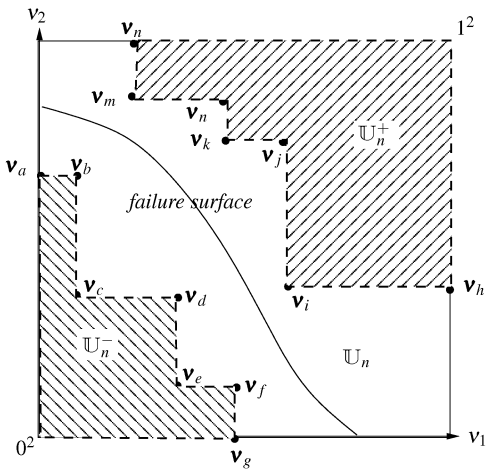


Figure 1: Dimension $d = 2$. Les points $\{0^2, \boldsymbol{\nu}_a, \boldsymbol{\nu}_b, \boldsymbol{\nu}_c, \boldsymbol{\nu}_d, \boldsymbol{\nu}_e, \boldsymbol{\nu}_f, \boldsymbol{\nu}_g\}$ sont situés en-dessous de la surface $G(\boldsymbol{\nu}) = 0$ et sont les sommets de $\mathbb{U}_n^-$. Les points $\{\boldsymbol{\nu}_h, \boldsymbol{\nu}_i, \boldsymbol{\nu}_j, \boldsymbol{\nu}_k, \boldsymbol{\nu}_l, \boldsymbol{\nu}_m, \boldsymbol{\nu}_n, 1^2\}$ sont situés au-dessus de la surface $G(\boldsymbol{\nu}) = 0$ et sont les sommets de $\mathbb{U}_n^+$.

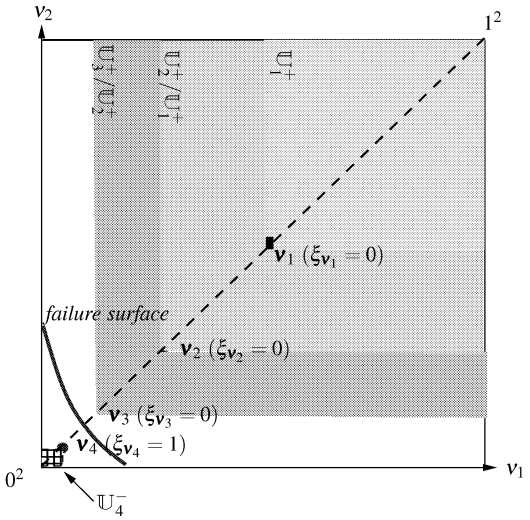


Figure 2: Plan d'expérience déterministe, dichotomique selon l'hyperplan diagonal séparateur, arrêté après 4 étapes.

consistant. Le calcul de ce dernier requiert au moins $n = 100/p_f$ appels à G (pour une erreur $\sim 10\%$ sur p_f ; cf. Lemaire et Pendola 2006) ce qui est peu réaliste en pratique. Il présente de plus le défaut d'être non-robuste, au sens où son coefficient de variation $\mathrm{CV}[\hat{p_f}_n] \to \infty$ quand $p_f \to 0$.

Dans la suite, deux estimateurs et leurs propriétés sont considérés. Jusqu'à la fin du texte, on supposera réinitialiser la notation $(\mathbb{U}_0^+, \mathbb{U}_0^-, p_0^+, p_0^-)$ après $N-1$ étapes déterministes où $N \geq k_0 + 1$. À partir de l'étape N, le plan d'expérience est choisi stochastique et on note $\mathscr{F}_n$ la $\sigma-$algèbre générée par la suite des n simulations $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$.

3 - Un estimateur par maximum de vraisemblance de données dépendantes

Dans un premier temps, on suppose que $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{n+1}$ sont successivement simulés uniformément dans la série de sous-espaces imbriqués $\mathbb{U}_0, \ldots, U_n$. L'occurence d'une *signature* $\xi_{\boldsymbol{\nu}_k} = \mathbb{1}_{\{G(\boldsymbol{\nu}_k) \leq 0\}}$ non-nulle suit une loi de Bernoulli $\mathscr{B}(\gamma_k)$, conditionnellement à $\mathscr{F}_{k-1}$, où

$$\begin{aligned} \gamma_k &= P\left(G(\boldsymbol{\nu}) \leq 0 | \boldsymbol{\nu} \in \mathbb{U}_{k-1}\right), \\ &= \frac{P\left(G(\boldsymbol{\nu}) \leq 0\right) - P\left(G(\boldsymbol{\nu}) \leq 0 | \boldsymbol{\nu} \in \mathbb{U}_{k-1}^-\right) P\left(\boldsymbol{\nu} \in \mathbb{U}_{k-1}^-\right)}{P\left(\boldsymbol{\nu} \in \mathbb{U}_{k-1}\right)} \end{aligned}$$

de par la règle de Bayes, d'où $\gamma_k = (p_f - p_{k-1}^-)/(p_{k-1}^+ - p_{k-1}^-)$. Après n étapes, toute l'information sur p_f est apportée par la vraisemblance de données dépendantes $L_n(p_f) = L_n(p_f|\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n)$ définie par le produit des conditionnements

$$L_n(p_f) = \prod_{k=1}^{n} \left(\frac{p_f - p_{k-1}^-}{p_{k-1}^+ - p_{k-1}^-} \right)^{\xi_{\boldsymbol{\nu}_k}} \left(\frac{p_{k-1}^+ - p_f}{p_{k-1}^+ - p_{k-1}^-} \right)^{1-\xi_{\boldsymbol{\nu}_k}}.$$

On pose $\ell_n(p_f) = \log L_n(p_f)$. On peut montrer qu'il existe une unique solution (EMV) $\hat{p_f}_n$ dans $[p_{n-1}^-, p_{n-1}^+]$ à l'équation de vraisemblance $\ell_n'(p_f) = 0$, semi-explicitement définie par

$$\hat{p_f}_n = \frac{\sum_{k=1}^{n} \tilde{\omega}_k\left(\hat{p_f}_n\right) p_k}{\sum_{k=1}^{n} \tilde{\omega}_k\left(\hat{p_f}_n\right)}, \tag{1}$$

avec $\tilde{\omega}_k(x) = \left[\left(x - p_{k-1}^-\right)\left(p_{k-1}^+ - x\right)\right]^{-1}$ et $p_k = p_{k-1}^- + \left(p_{k-1}^+ - p_{k-1}^-\right)\xi_{\boldsymbol{\nu}_k}$. L'étude des propriétés de cet estimateur (Bousquet 2010) nécessite d'utiliser la martingalité du processus $(\mathscr{F}_n)_n-$adapté associé au score $\{\ell_n'(p_f)\}$, et l'on peut ainsi obtenir la consistance faible (convergence en probabilité) ainsi que la normalité, la robustesse et une parcimonie asymptotiques sous des conditions modérées bien que difficiles à vérifier en pratique :

(i) la surface de défaillance est $\mathscr{C}^1$ sur un sous-ensemble mesurable non-vide de $\mathbb{U}$;

(ii) $\exists\, \delta \in [0,1)$ tel que $\sum_{k=1}^{n}(\tilde{\omega}_k(p_f) - \mathrm{E}_{\mathscr{F}_k}\left[\tilde{\omega}_k(p_f)\right]) = \mathrm{o}(n^{1-\delta})$;

(iii) $(p_n^+ - p_f)/(p_f - p_n^-) \xrightarrow[n\to\infty]{I\!P} 1$.

Alors $\frac{1}{\sigma_n}\left(\hat{p_f}_n - p_f\right) \xrightarrow{\mathscr{L}} \mathscr{N}(0,1)$ où $\sigma_n^2 < p_f(1-p_f)/n$ atteint la borne de Cramer-Rao (et constitue donc un optimum en variance pour tout estimateur sans biais fondé sur un échantillonnage progressif uniforme). Des conditions supplémentaires, mais elles aussi raisonnables, sont nécessaires pour estimer cette variance et obtenir la consistance presque sûre de $\hat{p_f}_n$.

Si cet estimateur est très simple à calculer et présente une forte réduction de variance vis-à-vis de Monte Carlo, il souffre d'un biais conservatif (positif lorsque p_f est petit) qui n'est pas forcément gênant dans un cadre fiabiliste général. Ce biais peut cependant devenir trop conservatif lorsque la dimension augmente. Par ailleurs une exploration uniforme de l'espace apparaît quelque peu naïve et les résultats obtenus sont tous asymptotiques. On cherche donc par la suite à produire une classe d'estimateur sans biais, robuste et de variance non-asymptotique significativement plus faible que celle de Monte Carlo.

4 - Une classe d'estimateurs par échantillonnage d'importance pondéré

On note $f_{k-1}(\boldsymbol{\nu})$ n'importe quelle densité d'importance de support $\mathbb{U}_{k-1}$, selon laquelle on simule le point $\boldsymbol{\nu}_k$ ajouté au plan d'expérience à l'itération k. Soit

$$\tilde{p}_k \;=\; p_{k-1}^- + \frac{\xi_{\boldsymbol{\nu}_k}}{f_{k-1}(\boldsymbol{\nu}_k)},$$

une généralisation de l'estimateur local p_k impliqué dans (1). On peut aisément prouver qu'il est sans biais. Dans ce cas, une généralisation non biaisée de l'EMV est l'estimateur pondéré WISE (*weighted importance sampling estimator*)

$$\breve{p}_{f_n} \;=\; \frac{1}{n}\sum_{k=1}^{n} \omega_k \tilde{p}_k, \tag{2}$$

où les ω_k sont des poids déterministes dans $[0,n]$, de somme égale à n. La variance de cet estimateur est $V_n^{WISE} = V_n^{\mathrm{U}}(p_f) + V_n^{\mathrm{F}}(p_f)$ où

$$\begin{aligned} V_n^{\mathrm{U}}(p_f) &= V_n^{MC}(p_f)\sum_{k=1}^{n}\frac{\omega_k^2}{n}\,(1-c_{k-1})\,, \\ V_n^{\mathrm{F}}(p_f) &= \sum_{k=1}^{n}\frac{\omega_k^2}{n^2}\mathrm{E}\left[\int_{\mathbb{U}_{k-1}}\frac{\mathbb{1}_{\{G(\boldsymbol{\nu})\leq 0\}}}{f_{k-1}(\boldsymbol{\nu})}\,d\boldsymbol{\nu} - \frac{(p_f - p_{k-1}^-)}{(p_{n-1}^+ - p_{n-1}^-)^{-1}}\right] \end{aligned}$$

avec $V_n^{\mathrm{F}}(p_f) = 0$ si tous les f_{k-1} sont supposés uniformes, les $\{c_k\}$ étant définis par $c_0 = 0$ et $\forall\, k > 1$,

$$c_k \;=\; \mathrm{E}_{\mathscr{F}_k}\left[\frac{p_k^-}{p_f} + \frac{1-p_k^+}{1-p_f} - \frac{p_k^-(1-p_k^+)}{p_f(1-p_f)}\right], \tag{3}$$

qui croît de 0 à 1 quand $k \to \infty$. Avec peu de contraintes sur f_k (en particulier f_k peut être naïvement uniforme), il y a toujours réduction de la variance (vis-à-vis d'un estimateur classique de Monte Carlo) et robustesse non-asymptotiques. La minimisation de la variance passe par la calibration des poids selon la règle $\omega_k^* \propto (1-d_{k-1})^{-1} / \sum_{j=1}^{n}(1-d_{j-1})^{-1}$ où $d_0 = 0$, $d_k = c_k - \mathrm{E}[b_k(f_k)]/(p_f(1-p_f))\ \forall k > 1$ et

$$b_k(f_k) \;=\; \int_{\mathbb{U}_k}\frac{\mathbb{1}_{\{G(\boldsymbol{\nu})\leq 0\}}}{f_k(\boldsymbol{\nu})}\,d\boldsymbol{\nu} - \frac{(p_f - p_k^-)}{(p_n^+ - p_n^-)^{-1}}.$$

Il faut noter qu'un choix particulier de f_{k-1} peut mener à une variance nulle. Il s'agit d'un résultat usuel en échantillonnage d'importance. Ce choix requiert de connaître p_f et la surface de défaillance $\mathbb{U}_l = \{\boldsymbol{\nu} \in \mathbb{U},\ G(\boldsymbol{\nu}) = 0\}$, ce qui est précisément le problème.

On peut tenter de s'approcher de ce choix optimal en simulant un point $\boldsymbol{\nu}_k$ tel que l'élément de variance conditionnelle $1 - c_k = \mathrm{E}[(p_f - p_k^-)(p_k^+ - p_f)]$ soit faible, soit en définissant f_{k-1} comme une densité unimodale, de variance très faible, dont le mode minimise le critère

$$C_{k-1}(\boldsymbol{\nu}) \quad = \quad \left(\breve{p}_{f_{k-1}} - p_k^-(\boldsymbol{\nu})\right)\left(p_k^+(\boldsymbol{\nu}) - \breve{p}_{f_{k-1}}\right)$$

Le mode de f_{k-1} est alors le point de $\mathbb{U}_{k-1}$ qui permet à une stratégie déterministe d'ôter le plus de "volume", ce dernier mot pouvant être défini come une mesure moyenne de l'écart prédictif entre les bornes. On remarque cependant que nécessairement, un tel critère ne peut être mis en oeuvre que si l'on dispose, pas à pas, d'un estimateur de la surface $\mathbb{U}_l$. Celui-ci est un *classifieur supervisé* d'un problème entièrement séparable qui respecte les contraintes suivantes : monotonie décroissante et volume sous la surface égale à l'estimateur courant de p_f. Les vecteurs de supervision peuvent être simulés uniformément dans les zones dominées sans limite de taille. La résistance au fléau de la dimension et la flexibilité requise favorise le choix (Hurtado 2004) de Perceptrons Multi-Couches ou de *Support Vector Machines* . Un tel classifieur permettrait de plus de mener des expérimentations bootstrap pour tenter de débiaiser l'EMV.

Ce problème n'est actuellement pas résolu et demande d'importants approfondissements. Mais en supposant que de telles lois d'importance, finalement réductibles à des multinormales tronquées de très faible variance, puissent être ainsi construites, quelques résultats importants peuvent en être déduits :

- de très faibles variance d'importance peuvent être calibrées de façon à ce que $\breve{p}_{f_n} \in [p_n^-, p_n^+]$ avec une très forte probabilité ;
- elles peuvent impliquer un quasi-déterminisme formalisé par $\mathrm{Var}[p_n^{-+}] \to 0$;
- ce dernier résultat assure la consistance presque sûre et la normalité asymptotique de la famille d'estimateurs.

Quelques expérimentations préparatoires en faible dimension ont montré que la variance de $\breve{p}_{f_n}$ pouvait se révéler numériquement au moins égale à la variance asymptotique de l'EMV. Celle-ci étant optimale pour l'échantillonnage uniforme naïf, ce résultat est prometteur, et l'on peut espérer obtenir des variances d'estimation encore plus faible en pratique si une résolution numérique au problème de classification posé ci-dessus peut être faite. Ainsi, de nombreux travaux théoriques et appliqués sont encore nécessaires pour proposer une construction automatisée de ces lois d'importance, la classification sous contrainte étant le premier d'entre eux.

Bibliographie

[1] de Rocquigny E. (2009). Structural Reliability under monotony: A review of Properties of FORM and associated simulation methods and a new class of monotonous reliability methods (MRM). *Structural Safety*, 31, 363–374.

[2] Limbourg, Ph., de Rocquigny, E., Andrianov, G. (2010). Accelerated uncertainty propagation in two-level probabilistic studies under monotony. *Reliab. Eng. Syst. Safety*, 95, 998–1010.
[3] Bousquet, N. (2010). Bounding and estimating an exceedance probability in output from monotonous time-consuming computer codes. ArXiv:1012.1042 (soumis).
[4] Lemaire, M., Pendola, M. (2006). PHIMECA-SOFT. *Structural Safety*, 28, 130–149.
[5] Hurtado, J.E. (2004). An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory. *Structural Safety*, 26, 271–293.

Hiérarchisation des sources d'incertitude vis à vis d'une probabilité de dépassement de seuil - Une méthode basée sur la pondération des lois

Paul Lemaître & Aurélie Arnaud

EDF R&D, 6 Quai Watier, 78401 Chatou

Résumé

Lors de la réalisation d'une analyse de sensibilité d'un modèle numérique, la plupart des méthodes existantes sont pensées pour le cas où la quantité d'intérêt est une variance. Les quelques méthodes disponibles quand la quantité d'intérêt est une probabilité sont soit basées sur des hypothèses restrictives, soit nécessitent un grand nombre d'évaluations. Cet article présente une méthode basée sur le tirage d'importance pour estimer l'impact de la modification d'un paramètre d'une loi d'entrée sur la probabilité de dépassement d'un seuil par la sortie d'un modèle. Cet outil s'avère être un complément intéressant pour une analyse de sensibilité quand le code est coûteux en temps de calcul, car il ne se base que sur des simulations déjà effectuées.

Mots-clés : Analyse de sensibilité ; incertitude ; fiabilité ; échantillonnage préférentiel

Abstract

When sensitivity analysis is performed on a numerical model, most methods are based on variance decomposition between the inputs and the output. However, the ones available when the quantity of interest is a probability use restrictive hypothesis or require a high number of function calls. This paper presents a method based on importance sampling to estimate the impact of modification of an input distribution parameter on the probability of a threshold crossing by the model output. This method appears useful for sensitivity analysis when the model computation is time-consuming, for it only uses previously-done simulations.

Keywords : Sensitivity analysis ; uncertainty ; reliability ; importance sampling

1 Introduction

Ce papier se place dans le cadre global d'études basées sur un modèle numérique simulant des phénomènes physiques, et dans le cadre plus précis de l'analyse de sensibilité (AS) de ce modèle numérique. Le but principal en AS est de hiérarchiser les variables d'entrée qui influent le plus sur la sortie du modèle comme décrit dans Saltelli et al. (2000). Il existe des techniques rodées qui décomposent la variance de la sortie en parts de variance due

à chaque entrée ou combinaison d'entrée, la plus célèbre étant les indices dits de Sobol (1993). Notre problématique concerne la fiabilité donc l'intérêt est porté sur la probabilité de dépassement d'un seuil (ou défaillance) par la sortie. Nous sommes intéressés par la sensibilité aux paramètres des lois d'entrée. Une méthode d'AS est présentée dans le cas d'un calcul de fiabilité dite "de pondération des lois" basée sur le tirage d'importance. C'est une méthode largement inspirée des travaux de Beckman & McKay (1987) et de ceux de Hesterberg (1996).

2 Proposition d'indices de sensibilité

Soit G la fonction d'état limite, définie dans l'espace des paramètres du modèle et soit X le vecteur aléatoire de $\mathbb{R}^d$ de densité de probabilité f_X des entrées du modèle. Dans la suite de l'article, les composantes du vecteur X sont considérées indépendantes deux à deux. Le domaine de défaillance est défini par $D_f = \left\{G(x) \leq 0, x \in \mathbb{R}^d\right\}$. Le but est d'estimer

$$P = \int 1_{\{G(x)\leq 0\}} f_X(x)dx = \mathbb{E}[1_{\{G(X)\leq 0\}}]$$

Comme les méthodes d'intégration directes ne sont pas applicables sauf dans des cas très particuliers, les pratiques de type Monte-Carlo (MC) sont les plus répandues. L'estimateur naïf MC est :

$$\widehat{P} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{G(x^i)\leq 0\}}$$

où les x^i sont N vecteurs aléatoires i.i.d. simulés selon f_X. C'est un estimateur sans biais de la quantité d'intérêt, dont il est possible de contrôler la précision via sa variance et l'expression d'un intervalle de confiance.

On peut vouloir calculer la probabilité de défaillance en supposant que la loi f_X des entrées a changé pour $f_{\tilde{X}}$. Dans ce cas la quantité à estimer est :

$$\widetilde{P} = \int 1_{\{G(x)\leq 0\}} f_{\tilde{X}}(x)dx$$

Si N échantillons i.i.d x^i ont été simulés selon f_X, $\widetilde{P}$ peut être estimé par :

$$\widetilde{P_N} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{G(x^i)\leq 0\}} \frac{f_{\tilde{X}}(x^i)}{f_X(x^i)}$$

La façon de procéder présentée ici est inversée par rapport au tirage d'importance classique : on estime P via la méthode MC classique pour ensuite en déduire une estimation

de $\widetilde{P}$ en utilisant uniquement les tirages déjà réalisés. Il faut cependant que le support de la densité modifiée $f_{\tilde{X}}$ soit contenu dans le support de f_X. Le but de notre AS est de hiérarchiser l'influence d'une ou de plusieurs variables aléatoires sur la probabilité de défaillance en effectuant le moins d'appels possible à la fonction G, supposée coûteuse. Ainsi seuls les tirages issus des calculs ayant servi à estimer la probabilité de défaillance peuvent être utilisés.

On se place dans le U-espace (ou espace gaussien), un espace dans lequel toutes les variables suivent une loi normale centrée réduite, ces variables étant indépendantes deux à deux. Soit F_{X_i} la fonction de répartition de X_i et Φ la fonction de répartition de la loi normale centrée réduite. Soit le réel δ, la variable aléatoire $X_{i\delta}$ de fonction de répartition $F_{X_{i\delta}}$ est définie pour tout x appartenant au support de F_{X_i} par :

$$\Phi^{-1}(F_{X_{i\delta}}(x)) = \Phi^{-1}(F_{X_i}(x)) - \delta$$

Ce changement de loi revient à translater d'un facteur δ la loi de X_i dans le U-espace où l'AS sera effectuée. La méthode proposée étudie donc l'influence d'une modification d'un paramètre (la moyenne) de la loi d'une variable d'entrée.

$P_{i,\delta}$ est définie comme $\mathbb{P}(G(X_1, .., X_{i\delta}, ..., X_d) \leq 0)$. L'indice de sensibilité à la modification de loi (dans le cas où P est différente de 0) est défini par :

$$S^P_{i,\delta} = \frac{P_{i,\delta} - P}{P}$$

C'est un indice sans dimension qui est positif lorsque la modification de la loi de X_i augmente la probabilité de défaillance et négatif sinon. C'est également le pourcentage de variation induit par le changement de moyenne. Les modifications conjointes pour plusieurs variables peuvent être étudiées, en modifiant les variables $i_1, ..., i_k$ de facteurs $\delta_1, ...\delta_k$. La représentation de tels indices est plus ardue et fera l'objet de recherche future.

On présente ici la méthodologie à suivre.

Propagation de l'incertitude Cette première étape n'est pas spécifique à la méthode de hiérarchisation des sources d'incertitudes proposée. Il s'agit d'estimer la probabilité de défaillance P comme par exemple avec l'estimateur naïf MC rappelé plus haut (on peut aussi utiliser d'autres méthodes comme par exemple de type quasi-MC).

Transformation de l'espace Si l'espace d'entrée n'est pas le U-espace, une transformation isoprobabiliste est nécessaire. La transformation de Rosenblatt est conseillée par Lemaire (2005), elle sera notée T. Si X est le vecteur des variables d'entrée $(X_1, ..., X_d)$, alors $U = (U_1, .., U_d) = T(X)$ où les U_i, $i = 1, ..., d$ sont des variables aléatoires normales centrées réduites indépendantes deux à deux.

Estimation des indices Cette étape consiste à estimer $P_{i,\delta}$. Soit $g(U) = G(T^{-1}(U))$, x un échantillon de X, $u = T(x)$, $\mu_{i\delta}$ la loi jointe du vecteur $(U_1, ..., U_{i\delta}, ..., U_d)$, μ la loi jointe du vecteur $U = (U_1, ..., U_i, ..., U_d)$, φ la densité de la variable U_i et φ_δ la densité de la variable $U_{i\delta}$.

$$\mu_{i\delta}(u) = \varphi_\delta(u_i).\prod_{k=1,k\neq i}^{d}\varphi(u_k) \qquad \text{et} \qquad \mu(u) = \prod_{k=1}^{d}\varphi(u_k) = \varphi(u_i).\prod_{k=1,k\neq i}^{d}\varphi(u_k)$$

Donc

$$P_{i,\delta} = \int 1_{\{g(u)\leq 0\}}\mu_{i\delta}(u)du = \int 1_{\{g(u)\leq 0\}}\frac{\varphi_\delta(u_i)}{\varphi(u_i)}\mu(u)du = \mathbb{E}_\mu[1_{\{g(U)\leq 0\}}\frac{\varphi_\delta(U_i)}{\varphi(U_i)}]$$

Les estimateurs proposés sont :

$$\widehat{P^P_{i,\delta}} = \frac{1}{N}\sum_{j=1}^{N}1_{\{g(u^j)\leq 0\}}\frac{\varphi_\delta(u_i^j)}{\varphi(u_i^j)} \qquad \text{et} \qquad \widehat{S^P_{i,\delta}} = \frac{\widehat{P^P_{i,\delta}} - \widehat{P}}{\widehat{P}}$$

3 Application

Dans cette section, nous allons metttre en application des AS sur des fonctions jouets à l'aide des indices proposés dans la section précédente. A noter que dans certains cas réels il n'y a pas d'expression analytique pour G (comme par exemple des codes de calculs thermo-hydrauliques basés sur les éléments finis).

Le premier cas d'études est sans transformation isoprobabiliste, c'est à dire que la fonction de défaillance est définie dans le U-espace. Pour la fonction de défaillance choisie, une variable est muette, une variable a un effet linéaire et une autre quadratique. Elle est définie par :

$$G(X) = 30X_2 + 10(X_3 - 3)^2 + 40$$

où $X = (X_1, X_2, X_3)$. Cette probabilité est estimée par MC avec $N = 10^5$. Le nuage de points associé n'est pas tracé car la formule est triviale. La probabilité de dépassement est estimée à $\widehat{P} = 0,00387$. La méthode proposée est exécutée, les indices de sensibilité $\widehat{S^P_{i,\delta}}$ associés aux $P_{i,\delta}$ pour $i = 1, 2, 3$ sont tracés en figure 1. La hiérarchisation des variables est donc, par ordre décroissant d'influence, X_2, X_3, X_1. L'influence de X_1 est nulle. Une augmentation de la moyenne de la loi d'entrée de la variable X_2 diminue la probabilité de défaillance (-2%) alors qu'une diminution de sa moyenne augmente la probabilité (jusque à 24%). Une diminution de la moyenne pour la variable X_3 diminue la probabilité (-2%) et une augmentation de sa moyenne augmente la probabilité (jusqu'à 10%).

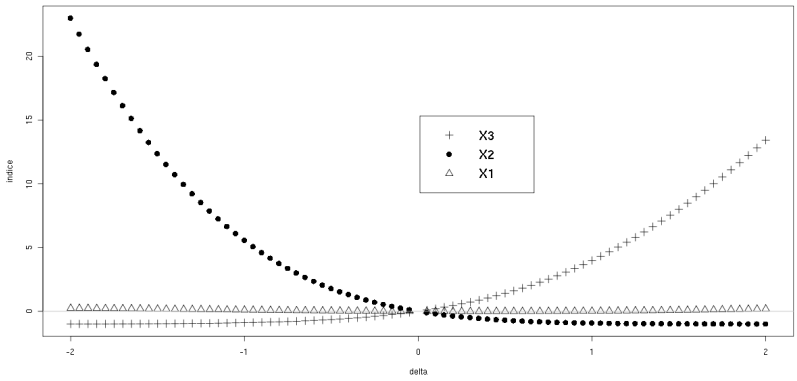


Figure 1: Indices de sensibilité - Premier cas

Le cas suivant est le cas d'une fonction classique en analyse de sensibilité, la fonction d'Ishigami. Cette fonction est définie pour $[-\pi;\pi]^3 \to \mathbb{R}$ et par :

$$G(X) = \sin(X_1) + 7.\sin^2(X_2) + 0,1.X_3^4.\sin(X_1)$$

Les X_i, $i = 1, 2, 3$, sont des lois $\mathcal{U}[-\pi;\pi]$. La quantité à estimer est $\mathbb{P}(G(X) \leq -7)$, ce seuil ayant été choisi arbitrairement afin d'avoir une probabilité de défaillance faible. L'estimation MC de P est effectuée dans l'espace d'entrée, le résultat est $\widehat{P} = 0,00612$. Les échantillons sont ensuite modifiés suivant la transformation de Rosenblatt afin de calculer les indices de sensibilité. Le nuage de point après transformation est tracé en figure 2.

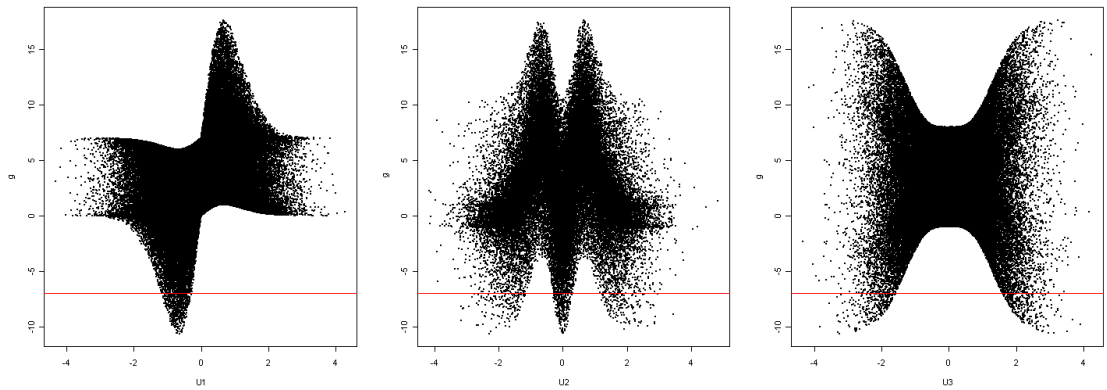


Figure 2: Nuage de points entre chaque entrée et la sortie du modèle - Second cas

Le caractère symétrique par rapport à 0 des variables U_2 et U_3 et l'aspect sinusoidal

de la variable U_1 sont visibles. Le graphique des indices estimés, c'est à dire les $\widehat{S^P_{i,\delta}}$ pour $i = 1, 2, 3$, est tracé en figure 3. La hiérarchisation des variables est, par ordre décroissant d'influence, U_3, U_2, U_1. La conclusion à tirer de l'étude de la figure 3 est que la variable U_3 a une influence prépondérante vis à vis de la probabilité de dépassement de seuil. On note que le résultat de l'AS pour la probabilité de dépassement de seuil diffère de celui de l'AS sur la variance (les indices de Sobol sont : $S_1 = 0,3138$, $S_2 = 0,4424$, $S_3 = 0$, $S_{1,3} = 0,2436$).

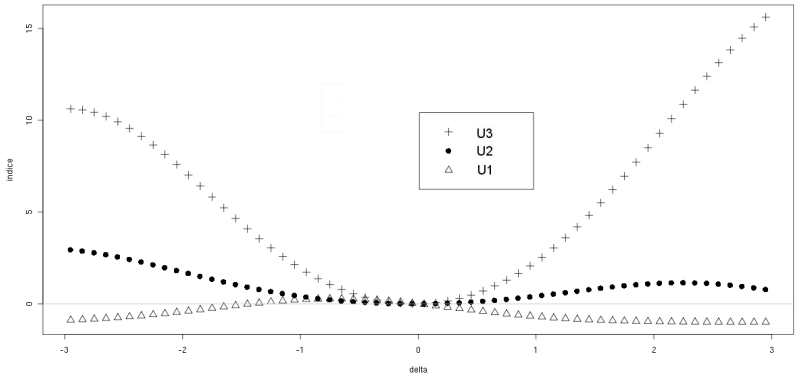


Figure 3: Indices de sensibilité - Second cas

4 Conclusion

Une méthodologie qui permet d'estimer l'impact relatif de la modification d'un paramètre d'une loi d'entrée du modèle sur la probabilité de défaillance a été présentée ici. La modification plus ou moins forte du paramètre représente l'incertitude, ses effets sur les valeurs de la probabilité de défaillance peuvent donc être quantifiés.

Bibliographie

[1] Saltelli, A. et al. (2000), Sensitivity analysis, Wiley New York.

[2] Beckman, R.J. et McKay, M.D (1987), Monte Carlo estimation under different distributions using the same simulation, *Technometrics*, 29, 2, 153–160.

[3] Hesterberg, TC. (1996), Estimates and confidence intervals for importance sampling sensitivity analysis, *Mathematical and Computer Modelling*, 23, 8-9, 79–85.

[4] Lemaire M. (2005), Fiabilité des structures : couplage mécano-fiabiliste statique, Hermès Science Publications.

[5] Sobol, IM. (1993), Sensitivity analysis for non-linear mathematical models, *Mathematical Modelling and Computational Experiment*, 1, 1, 407–414.

An Application of the AFT Model Based on the Generalized Weibull Distribution for Reliability Analysis of Redundant Systems.

N. Saaidia
Université Badji Mokhtar, Annaba, Algérie &
IMB, Université Victor Segalen Bordeaux 2, France.

M. Nikulin
IMB, Université Victor Segalen Bordeaux 2, France.

R. Tahir
IMB, Université Victor Segalen Bordeaux 2, France.

Résumé : En fiabilité, les distributions qui ont une fonction de hasard de forme $\cap$ ou unimodale ne sont pas trop nombreuses. Elles comprennent : log-normale, log-logistique, Weibull généralisée et Gaussienne inverse. Dans cette étude, nous illustrons l'application de la distribution de Weibull généralisée à l'analyse des systèmes redondants avec une unité principale et une unité de réserve. Les propriétés asymptotiques des estimateurs et des intervalles de confiance sont obtenues.

Mots-clés : Distribution de Weibull généralisée, Fiabilité, Modèle de Sedyakin, Systèmes redondants, Modèle de panne accélérée, Intervalle de confiance, Temps de panne, Maximum de vraisemblance, Estimation.

Abstract : In reliability, the distributions that have a hazard function of $\cap-$shape or unimodal are not too much. They include: log-normal, log-logistic, power generalized Weibull and inverse Gaussian distributions. In this study, we give the application of the generalized Weibull distribution in the analysis of redundant systems with one main unit and one stand-by unit. Asymptotic properties of the estimators and asymptotic confidence intervals are obtained.

Key-words : Generalized Weibull distribution, Reliability, Sedyakin's model, Redundant system, AFT model, Confidence interval, Failure time, Maximum Likelihood, Estimation.

1 Introduction

Let us consider redundant systems with one principal main unit operating in 'hot' and one stand-by unit operating in 'warm' conditions. The problem is to obtain confidence intervals for the cumulative distribution function of the system using failure data of two groups of units. We assume that switching from warm to hot does not cause shock or damage to units.

Suppose that in 'hot' conditions the failure time, the c.d.f, the survival function and the density of the main unit are T_1, F_1, S_1 and f_1 successively, and in 'warm' conditions the failure time, the c.d.f, the survival function and the density of the stand-by unit are T_i, F_2, S_2 and $f_2, i = 2, \cdots, m$. Based on the AFT and Sedykin's model Bagdonavičius et al. (2008a, 2008b, 2009, 2010) give mathematical formulation of "fluent switch on" and propose tests for verification of this hypothesis.
Suppose that the distribution of units operation in 'warm' and 'hot' conditions differ only in scale, i.e. $F_2(t) = F_1(rt)$ for all $t \geq 0$ and some $r > 0$, and we have the AFT model. In our case we suppose that the c.d.f. of units belong to the generalized Weibull family (see Bagdonavičius and Nikulin (2002)), i.e.

$$S_1(t) = 1 - F_1(t) = exp\left\{1 - (1 + (\frac{t}{\sigma})^{\nu})^{\frac{1}{\gamma}}\right\}, \quad t \geq 0, \quad \sigma, \nu, \gamma \in \mathbf{R}_+^*.$$

The corresponding density function is given by

$$f_1(t) = \frac{\nu}{\gamma\sigma^{\nu}} t^{\nu-1} \left(1 + (\frac{t}{\sigma})^{\nu}\right)^{\frac{1}{\gamma}-1} exp\left\{1 - (1 + (\frac{t}{\sigma})^{\nu})^{\frac{1}{\gamma}}\right\}, \quad t \geq 0.$$

In the case of $\cap-$shape hazard function, we have necessary $\gamma > \nu > 1$. The c.d.f. $K_2(t)$ of the system with one main unit and one stand-by unit is given recurrently by

$$K_2(t) = \int_0^t F_1(t + ry - y) dK_1(y), \;\; K_1(t) = F_1(t).$$

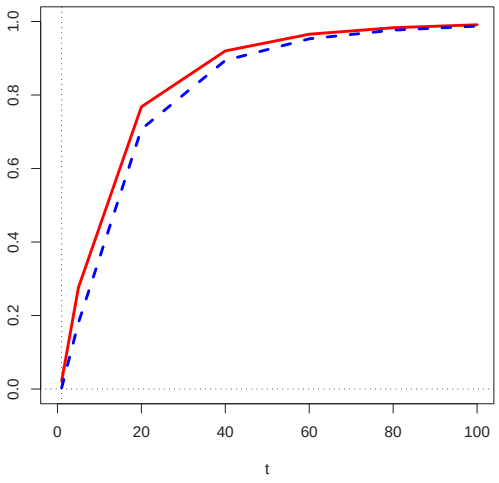


Fig. 1. Graphs of the trajectories of the parametric estimators $\hat{F}_1$ and $\hat{K}_2$ (power generalized Weibull distribution).

2 Point estimators of the K_j

Suppose that the following data are available:
a) complete ordered sample $T_{11}, \cdots, T_{1n_1}$ of size n_1, T_{1i} is the failure time of units tested in 'hot' condition;
b) complete ordered sample $T_{21}, \cdots, T_{2n_2}$ of size n_2, T_{2i} is the failure time of units tested in 'warm' condition.

Let $\gamma_1 = (r, \sigma, \nu, \gamma)^T$. The MLE's $\hat{\gamma}_1 = (\hat{r}, \hat{\sigma}, \hat{\nu}, \hat{\gamma})^T$ of the parameter γ_1 maximizes the loglikelihood function

$$\ell(r, \sigma, \nu, \gamma) = (\ln \nu - \ln \gamma - \nu \ln \sigma)\, n + \nu n_2 \ln r +$$

$$(\nu - 1) \sum_{i=1}^{n_1} \ln(T_{1i}) + \left(\frac{1}{\gamma} - 1\right) \sum_{i=1}^{n_1} \ln\left(1 + \left(\frac{T_{1i}}{\sigma}\right)^{\nu}\right) - \sum_{i=1}^{n_1} \left(1 + \left(\frac{T_{1i}}{\sigma}\right)^{\nu}\right)^{\frac{1}{\gamma}} +$$

$$(\nu - 1) \sum_{j=1}^{n_2} \ln(T_{2j}) + \left(\frac{1}{\gamma} - 1\right) \sum_{j=1}^{n_2} \ln\left(1 + \left(\frac{rT_{2j}}{\sigma}\right)^{\nu}\right) - \sum_{j=1}^{n_2} \left(1 + \left(\frac{rT_{2j}}{\sigma}\right)^{\nu}\right)^{\frac{1}{\gamma}},$$

where $n = n_1 + n_2$.
To find the estimator $\hat{\gamma}_1$ one can solve the system formed by equalizing the score functions to zero.

It is easy to calculate the second partial derivatives $\ddot{\ell}(\gamma_1)$ of the loglikelihood function.

The Fisher's information matrix is

$$I(\gamma_1) = -E\ddot{\ell}(\gamma_1),$$

and it may be replaced by $-\ddot{\ell}(\hat{\gamma_1})$.

The c.d.f. $K_2(t)$ is estimated by

$$\hat{K}_2(t) = \frac{\hat{\nu}}{\hat{\gamma}\hat{\sigma}^{\hat{\nu}}} \int_0^t \left(1 - e^{1 - \left(1 + \left(\frac{t + \hat{r}y - y}{\hat{\sigma}}\right)^{\hat{\nu}}\right)^{\frac{1}{\hat{\gamma}}}}\right) y^{\hat{\nu}-1} \left(1 + (\frac{y}{\hat{\sigma}})^{\hat{\nu}}\right)^{\frac{1}{\hat{\gamma}}-1} e^{1-(1+(\frac{y}{\hat{\sigma}})^{\hat{\nu}})^{\frac{1}{\hat{\gamma}}}} dy.$$

3 Asymptotic confidence interval for $K_2(t)$

Using the results from Bagdonavičius et al. (2008a, 2008b, 2009, 2010) we can construct the asymptotic $1 - \alpha$ confidence interval $(\underline{K}_2(t), \overline{K}_2(t))$ for $K_2(t)$, where

$$\underline{K}_2(t) = \left(1 + \frac{1 - \hat{K}_2(t)}{\hat{K}_2(t)} exp\left\{\frac{\hat{\sigma}_{\hat{K}_2} z_{1-\alpha/2}}{\sqrt{\hat{K}_2(t)(1 - \hat{K}_2(t))}}\right\}\right)^{-1},$$

$$\overline{K}_2(t) = \left(1 + \frac{1-\hat{K}_2(t)}{\hat{K}_2(t)} exp\left\{-\frac{\hat{\sigma}_{\hat{K}_2} z_{1-\alpha/2}}{\sqrt{\hat{K}_2(t)(1-\hat{K}_2(t))}}\right\}\right)^{-1},$$

where

$$\hat{\sigma}^2_{\hat{K}_2(t)} = C_2^T(t,\hat{\gamma}_1) I^{-1}(\hat{\gamma}_1) C_2(t,\hat{\gamma}_1),$$

$$C_2(t,\gamma_1) = (C_{21}(t,\gamma_1), C_{22}(t,\gamma_1), C_{23}(t,\gamma_1), C_{24}(t,\gamma_1))^T,$$

$$C_{21}(t,\gamma_1) = \int_0^t \frac{\partial F_1}{\partial r}(t+ry-y;\sigma,\nu,\gamma) dF_1(y;\sigma,\nu,\gamma),$$

$$C_{22}(t,\gamma_1) = \int_0^t \frac{\partial F_1}{\partial \sigma}(t+ry-y;\sigma,\nu,\gamma) dF_1(y,m,\sigma,\nu,\gamma) + F_1(t+ry-y;\sigma,\nu,\gamma) d(\frac{\partial F_1}{\partial \sigma}(y;\sigma,\nu,\gamma)),$$

$$C_{23}(t,\gamma_1) = \int_0^t \frac{\partial F_1}{\partial \nu}(t+ry-y;\sigma,\nu,\gamma) dF_1(y,m,\sigma,\nu,\gamma) + F_1(t+ry-y;\sigma,\nu,\gamma) d(\frac{\partial F_1}{\partial \nu}(y;\sigma,\nu,\gamma)),$$

$$C_{24}(t,\gamma_1) = \int_0^t \frac{\partial F_1}{\partial \gamma}(t+ry-y;\sigma,\nu,\gamma) dF_1(y,m,\sigma,\nu,\gamma) + F_1(t+ry-y;\sigma,\nu,\gamma) d(\frac{\partial F_1}{\partial \gamma}(y;\sigma,\nu,\gamma)).$$

Remark. Using the results of Bagdonavičius et al. (2010) it is easy to estimate the parameter γ_1 when the data are censored.

Suppose that the following data are available:
a) right censored sample

$$(X_{11},\delta_{11})^T, \cdots, (X_{1n_1},\delta_{1n_1})^T$$

of size n_1, where $X_{1i} = T_{1i} \wedge C_{1i}, \quad \delta_{1i} = 1_{\{T_{1i} \le C_{1i}\}}$;

b) right censored sample

$$(X_{21},\delta_{21})^T, \cdots, (X_{2n_2},\delta_{2n_2})^T$$

of size n_2, where $X_{2j} = T_{2j} \wedge C_{2j}, \quad \delta_{2j} = 1_{\{T_{2j} \le C_{2j}\}}$.

Let $m_1 = \sum_{i=1}^{n_1} \delta_{1i}, \quad m_2 = \sum_{j=1}^{n_2} \delta_{2j}$.

The loglikelihood function is given by

$$\ell(r,\sigma,\nu,\gamma) = \sum_{i=1}^{n_1} \delta_{1i} \ln f_1(X_{1i};\theta) + \sum_{i=1}^{n_1} (1-\delta_{1i}) \ln S_1(X_{1i};\theta) + m_2 \ln r +$$

Tab.1. Confidence level for samples with $n_1 = n_2 = 100$,
power generalized Weibull distribution

t	1	5	20	40	60	80	100
$K_2(t)$	0.0034	0.18036	0.7070	0.8941	0.9529	0.9765	0.9874
C.L. (%)	78.76	79.08	80.54	78.32	81.39	78.89	82.21

$$\sum_{j=1}^{n_2} \delta_{2j} \ln f_1(rX_{2j};\theta) + \sum_{j=1}^{n_2} (1-\delta_{2j}) \ln S_1(rX_{2j};\theta).$$

We can also write

$$\ell(r,\sigma,\nu,\gamma) = (\ln\nu - \ln\gamma - \nu\ln\sigma)(m_1+m_2) + \nu m_2 \ln r +$$

$$(\nu-1)\sum_{i=1}^{n_1}\delta_{1i}\ln(X_{1i}) + \left(\frac{1}{\gamma}-1\right)\sum_{i=1}^{n_1}\delta_{1i}\ln\left(1+\left(\frac{X_{1i}}{\sigma}\right)^{\nu}\right) - \sum_{i=1}^{n_1}\left(1+\left(\frac{X_{1i}}{\sigma}\right)^{\nu}\right)^{\frac{1}{\gamma}} +$$

$$(\nu-1)\sum_{j=1}^{n_2}\delta_{21j}\ln(X_{2j}) + \left(\frac{1}{\gamma}-1\right)\sum_{j=1}^{n_2}\delta_{2j}\ln\left(1+\left(\frac{rX_{2j}}{\sigma}\right)^{\nu}\right) - \sum_{j=1}^{n_2}\left(1+\left(\frac{rX_{2j}}{\sigma}\right)^{\nu}\right)^{\frac{1}{\gamma}},$$

and the confidence interval for $K_2(t)$ is calculated as previously .

4 Simulation study

Let us consider the case of complete sample of size $n_1 = n_2 = 100$. Each sample is repeated 5000 times. We find by simulation the confidence levels of intervals using asymptotic formulas with $1-\alpha = 0.8$. We simulated failure times T_{1i} and T_{2j} from power generalized Weibull distribution with the parameters:

$$T_{1i} \sim PGW(\sigma_1,\nu_1,\gamma_1), \quad T_{2j} \sim PGW(\sigma_2,\nu_2,\gamma_2),$$

$$\sigma_1 = 4, \quad \sigma_2 = 4/5, \quad \nu_1 = \nu_2 = 2, \quad \gamma_1 = \gamma_2 = 4.$$

For various values of t the proportions of confidence interval (C.L.) realizations covering the true value of the distribution function $K_2(t)$ are given in table 1.

5 Conclusion

Here we considered the application of the power generalized Weibull family in redundant systems which is a very important family in reliability. Noting that for specified parameters the hazard rate function of this family is unimodal like the log-normal, log-logistic and inverse Gaussian distributions. It will be very important to make a comparative study with these four families and that is the subject of future work.

References

[1] Bagdonavičius V., Nikukin M. (2002) *Accelerated Life Models: Modeling and Statistical Analysis*, Chapman and Hall.

[2] Bagdonavičius V., Masiulaityle I., Nikulin M. (2008a) *Statistical analysis of redundant system with "warm" stand-by units*, Stochastics: An International Journal of Probability and Stochastical Processes, 80 #2-3, 115–128.

[3] Bagdonavičius V., Masiulaityle I., Nikulin M. (2008b) *Statistical analysis of redundant system with one stand-by unit*, In: *Mathematical methods in Survival analysis, reliability and quality of life* (Huber C., Limnios N., Balakrishnan N., Messbah M., Nikulin M. (Eds)), ISTE & Wiley: London, 189–202.

[4] Bagdonavičius V., Masiulaityle I., Nikulin M. (2009) *Asymptotic Properties of Redundant Systems Reliability Estimators*, In: *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis, and Finance* (Nikulin M., Limnios N., Balakrishnan N., Kahle W., Huber C. (Eds)), Birkhäuser Boston, 1st Edition, 293–310.

[5] Bagdonavičius V., Masiulaityle I., Nikulin M. (2010) *Parametric estimation of redundant system reliability from censored data*, In: *Mathematical and Statistical Methods in Reliability* (Balakrishnan N., Nikulin M., Rykov V. (Eds)), Springer Verlag: Boston, 51–64.

[6] Meeker W.Q., Escobar L.A. (1998) *Statistical Methods for reliability Data*, John Wiley and Sons, INC.

[7] Singpurwalla N.D. (2006) *Reliability and Risk*, J.Wiley, Chichester.

[8] Nikulin M.S., Saaidia N. (2009) *Inverse Gaussian family and its applications in reliability. Study by simulation.* Proceedings of the 6th St. Petersburg Workshop on Simulation, St. Petersburg, June 28-July 4. VVM comm. Ltd., St. Petersburg, V. 2, 657–661.

[9] Rykov V.V., Balakrishnan N., Nikulin M.S. (Eds)(2010) *Mathematical and Statistical Models and Methods in Reliability, Applications to Medicine, Finance, and Quality Control Series: Statistics for Industry and Technology*, Birkhäuser Boston.

Réflexions sur l'analyse d'incertitude dans un contexte industriel: information disponible et enjeux décisionnels

Merlin Keller[1] & Alberto Pasanisi[1] & Khouloud Ghorbel[1,2] & Eric Parent[2]
[1] EDF R&D [2] AgroParisTech

EDF-R&D, 6 quai Watier, 78401 Chatou

Résumé. L'analyse d'incertitude sert à quantifier l'incertitude affectant la valeur d'une quantité d'intérêt, caractéristique du fonctionnement d'un sytème physique, et liée à des enjeux décisionnels importants. La plupart des approches font appel à l'inférence statistique, et se divisent en trois grandes classes: les approches classiques, fournissant une estimation ponctuelle de la quantité d'intérêt; les procédures bayésiennes qui déduisent une valeur optimale de la grandeur d'intérêt d'une fonction de coût formalisant les enjeux décisionnels; et enfin, les approches purement descriptives, visant à décrire l'incertitude sur la variable d'intérêt. La pertinence de chaque approche varie en fonction des enjeux de l'analyse d'incertitude, et de la connaissance qu'en a l'analyste. Nous illustrons ce propos sur un jeu de données réelles de hauteurs et débits d'un cours d'eau.

keywords: *Bayesian methods; Industrial studies; Decision theory; Uncertainty analysis;*

Abstract. Uncertainty analysis aims at quantifying the uncertainty affecting the value of a quantity of interest, characteristic of a the functioning of a physical system, and related to important decisional issues. Most approaches rely on statistical inference, and are divided into three classes: classical approaches, providing a point estimate of the quantity of interest; Bayesian procedures that infer an optimal value of the quantity of interest from a cost function formalizing the decisional stakes, and finally purely descriptive approaches, that describe the uncertainty affecting the quantity of interest. The relevance of each approach varies depending on the goals of the uncertainty analysis, and the amount of data available to the analyst. We illustrate this on a real dataset of a water height and discharge measures.

mots-clés: *méthodes bayésiennes. ingénierie-industrie; théorie de la décision; Analyse d'incertitudes;*

1 Modèle déterministe

Nous considérons ici un modèle hydraulique simplifié de relation débit/hauteur pour un tronçon de rivière, résultant des équations de Saint-Venant en 1D sous hypothèse de stationnarité et de section rectangulaire très large. La fonction G se présente alors sous une forme analytique :

$$Z_c \;=\; Z_v + \left\{ Q \,/\, \left(BK_s \sqrt{(Z_m - Z_v)/L} \right) \right\}^{0.6}, \tag{1}$$

où Z_c représente la cote de la surface de la rivière en aval (en m), ou *hauteur de crue*, Z_m et Z_v les cotes du fond de la rivière en amont et en aval respectivement (en m), Q le débit de crue (en m^3/s), B la largeur du cours d'eau (en m), K_s le coefficient de Strickler, et L la longueur (en m) du tronçon considéré.

2 Probabilisation des entrées

Dans le modèle hydraulique (1), nous modélisons Q représentant le débit maximal annuel, par une variable aléatoire de loi de Gumbel $Gu(\mu, \rho)$, de fonction de répartition

$$F(t) = \exp\{-\exp[\rho(\mu - t)]\}. \tag{2}$$

La loi de la sortie Z_c, qui représente la hauteur d'eau maximal annuelle, est alors explicite :

$$H(t) \;=\; \exp\left(-\exp\left\{ \rho\left(\mu - BK_s \sqrt{\frac{Z_m - Z_v}{L}} (t - Z_v)^{5/3} \right) \right\} \right). \tag{3}$$

3 Quantité d'intérêt

Dans le modèle défini par les équations (1) et (2), nous considérons une quantité d'intérêt Φ, caractéristique de la loi H de Z_c, et qui résume les enjeux décisionnels de l'étude, selon la démarche décrite entre autres dans E. de Rocquigny (2006). Nous considérons deux définitions alternatives pour Φ :

- Le quantile $\Phi = q_\alpha = H^{-1}(\alpha)$, donné par la formule

$$q_\alpha \;=\; Z_v + \left\{ \left(\mu - \frac{1}{\rho} \log\log \frac{1}{\alpha} \right) \Big/ \left(K_s \, B \sqrt{\frac{Z_m - Z_v}{L}} \right) \right\}^{0.6}. \tag{4}$$

- La hauteur de digue optimale $\Phi = h_{opt} = \arg\min_d c(d\,; H)$, minimisant la fonction de coût (proposée dans Bernier (2003)):

$$c(d\,; H) \;=\; I_0 \times d + C_0 \times \mathbb{E}[\mathbf{1}_{Z_c > d} (Z_c - d)^2], \tag{5}$$

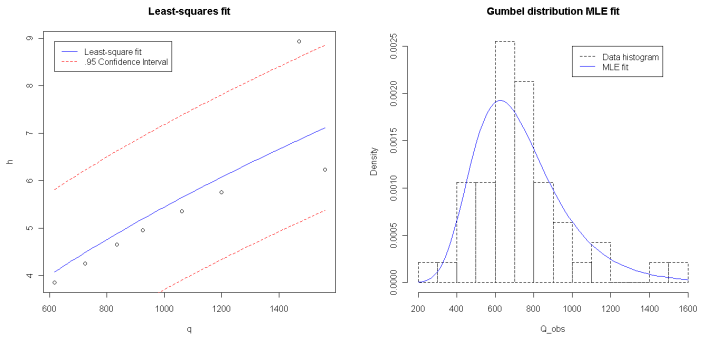


Figure 1: Gauche: Ajustement par moindre carrés du modèle (1) à $m = 8$ couples de mesures débit/hauteur. Droite: estimation par maximum de vraisemblance du modèle de Gumbel à partir de $n = 47$ mesures de débits maximaux annuels.

représentant le coût résultant de la construction d'une digue de protection fluviale d'altitude d (en m); I_0 représente un coût d'investissement marginal et C_0 un coût de dommage marginal.

4 Incertitude épistémique

Dans le modèle hydraulique (1), les grandeurs physiques représentées par les constantes Z_v, B, K_s et L, qui déterminent la transformation G, sont inconnues. Elles sont donc affectées d'une incertitude dite *epistémique*, c'est-à-dire dûe à une connaissance imparfaite.

Par opposition, le débit représenté par la variable Q, varie de manière imprévisible d'une année sur l'autre: il s'agit donc d'une grandeur incertaine par nature, puisqu'on ne pourra jamais la prédire exactement sa valeur. Enfin, la distribution F (2) de la variable Q est elle-même, comme G, affectée d'une incertitude épistémique.

Le couple $\Theta = (F, G)$ caractérise le fonctionnement du système physique considéré, encore appelé *état de la nature*. On le notera également $\Theta = (K_s, \sigma, \mu, \rho)$ car ces quatre paramètres définissent ici complétement F et G. De même, Θ définit complètement la quantité d'intérêt Φ décrite plus haut, que nous notons donc $\Phi = \Phi(\Theta)$.

5 Estimation classique

Dans un premier temps, nous avons choisi d'estimer la quantité d'intérêt $\Phi(\Theta)$ de faqn classique, c'est-à-dire en remplaçant la valeur inconnue de Θ par un estimateur ponctuel: $\widehat{\Phi} = \Phi(\Theta)$. Pour ce faire, nous disposons d'un jeu de données composé de $m = 8$ couples $(q_i, y_i)_{1 \leq i \leq m}$ de mesures débit/hauteur effectués sur un cours d'eau, ainsi que d'une série de $n = 47$ mesures $(Q_j)_{1 \leq j \leq n}$ de débits maximaux annuels effectués sur le même site.

Nous avons supposé que les valeurs q_i et Q_j de débits étaient observées sans erreurs, mais que les hauteurs d'eau y_i étaient mesurées avec une erreur additive, modélisée par un bruit blanc gaussien, d'écart-type σ inconnu. Nous avons alors ajusté le modèle (1) par moindre carrés ordinaires (MCO) aux couples (q_i, y_i), et estimé le modèle de Gumbel (2) par maximum de vraisemblance (MV) sur la base des Q_j. On obtient : (voir Figure 1)

$$\widehat{K_s}^{\text{MCO}} = 59.33; \qquad \widehat{\mu}^{\text{MV}} = 626.14; \qquad \widehat{\rho}^{\text{MV}} = 5.24 \times 10^{-3}.$$

Par substitution de ces valeurs dans les équations (3), (4) et (5), on obtient alors pour les définitions de Φ données plus haut:

$$\widehat{q}_{99\%} = Z_v + 6.96\, m \qquad \widehat{h}_{opt} = 8.18\, m.$$

la hauteur auteur optimale de digue étant calculée pour $I_0/C_0 = 1/1\,000$.

L'estimation classique décrite ci-dessus permet de calculer relativement simplement des estimateurs de n'importe quelle quantité d'intérêt voulue, et donne des résultats facilement interprétables. En revanche, nous avons négligé l'erreur commise en estimant Θ, et donc Φ. Se posent alors plusieurs questions:

- Peut-on quantifier l'erreur d'estimation $(\widehat{\Theta} - \Theta)$, ainsi que l'erreur résultante sur l'estimation de la quantité d'intérêt $(\widehat{\Phi} - \Phi)$?
- Quelles sont les conséquences de ces erreurs sur la décision que doit guider l'étude?

6 Estimation bayésienne

Dès lors que la définition de Φ s'appuie sur une fonction de coût, comme dans l'un des deux cas considérés ici, la théorie de l'estimation bayésienne, développée par exemple dans E. Parent, J. Bernier (2007), offre un cadre cohérent permettant de répondre aux deux questions posées plus haut. Étant donnée une loi $\pi(\Theta)$ sur les inconnues du modèle, la décision optimale au sens de Bayes correspond au minimum du coût intégré:

$$\widehat{\Phi}_{\pi}^{Bayes} \quad = \quad \arg\min_{d} \int c(d; H(\Theta))\pi(\Theta|D)\mathrm{d}\Theta, \tag{6}$$

où $\pi(\Theta|D) \propto p(D|\Theta)\pi(\Theta)$ donne formellement la loi de l'inconnue Θ conditionnellement aux données, $p(D|\Theta)$ étant la vraisemblance de celles-ci.

Adoptant cette démarche, nous avons utilisé les lois *a priori* suivantes sur les paramètres $\Theta = (K_s, \sigma, \mu, \rho)$ du modèle:

$$K_s \sim \mathcal{U}([10\,;100]); \qquad 1/\sigma^2 \sim \mathcal{E}(1); \qquad \mu \sim \mathcal{G}(1\,;500); \qquad \rho \sim \mathcal{G}(1\,;200),$$

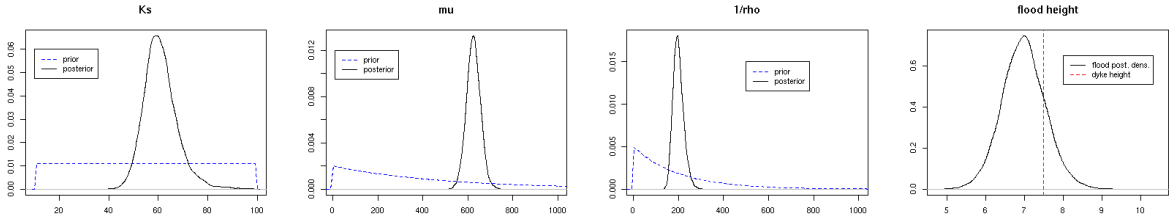


Figure 2: Distributions marginales *a posteriori* des paramètres du modèle hydraulique, approchées par un estimateur à noyau à partir d'un échantillon de 10 000 valeurs. De gauche à droite: K_s, μ, ρ, et $q_{99\%} - Z_v$.

où $\mathcal{G}(\alpha\,;\beta)$ est la distribution Gamma de paramètre de forme α et d'échelle inverse β.

A partir d'un échantillon de la loi *a posteriori* de Θ, il devient alors facile de résoudre numériquement le problème d'optimisation (6) Appliquée au problème de calcul d'une hauteur optimale de digue, pour $I_0/C_0 = 1/1\,000$, cett approche donne $\widehat{h}_{opt}^{Bayes} = 10.76\,m$.

Ce résultat est à comparer à celui obtenu dans la Section précédente: $\widehat{h}_{opt} = 8.18\,m$. De manière informelle, la prise en compte des incertitudes épistémiques sur les paramètre du modèles, se traduisant en incertitudes sur les coûts de dommage potentiels, entraîne dans le cas présent un relèvement de la hauteur optimale de digue de presque 3 m, et mène donc à une décision plus "prudente". Une telle observation se retrouve dans Bernier (2003) pour la même fonction de coût et un modèle similaire.

7 Approche descriptive

On se donne à présent pour but de vérifier que la hauteur de crue centennale $q_{99\%} - Z_v$ est inférieure à celle d'une digue fictive de hauteur $h = 7.5\,m$, ce qui s'interprète en disant qu'elle ne sera dépassée qu'une fois tout au plus dans le siècle à venir. Rappelons que l'estimation ponctuelle par maximum de vraisemblance donne: $\widehat{q}_{99\%} = Z_v + 6.96\,m$, ce qui suggère que le risque est bien en deça de la limite souhaitée. Cependant, étant donnée l'incertitude sur la valeur réelle de $q_{99\%}$, on peut légitimement se demander dans quelle mesure cette première réponse est fiable.

Continuant l'analyse bayésienne commencée dans la Section 6, nous avons calculé la distribution *a posteriori* du quantile $q_{99\%}$ (voir Figure 2, à droite). La probabilité *a posteriori* que ce quantile dépassait la hauteur de la digue est égale à: $\mathbb{P}[q_{99\%} > Z_v + h] = 0.17$, ce qui est loin d'être négligeable. Il peut alors sembler raisonnable d'envisager des travaux d'agrandissement de la digue pour la rendre plus fiable.

Ce cas précis illustre le danger qu'il y a à ne considérer que des estimateurs ponctuels sans tenir compte de l'incertitude affectant les grandeurs d'intérêt considérées, et la nécessité pour l'analyste de fournir le maximum d'information au commanditaire de l'étude afin de lui permettre une prise de décision la plus renseignée possible.

8 Discussion

En conclusion de cette réflexion, il nous semble important de mettre en avant les points suivants, qui devraient éclairer l'analyste dans le choix d'une méthodologie.

Dans un premier temps, il nous semble que le critère le plus important est la quantité d'information disponible. En effet, en situation d'information parfaite ou presque, c'est-à-dire lorsqu'on dispose d'assez de données pour estimer le modèle avec précision, on peut utiliser n'importe quelle méthode d'estimation ponctuelle consistante.

Si l'on ne dispose que de peu de données, il est impératif de prendre en compte les incertitudes épistémiques affectant les quantitées d'intérêt à évaluer. Un deuxième critère est alors l'accès à une fonction de coût formalisant les enjeux décisionnels de l'étude, auquel cas le choix d'une approche bayésienne est à privilégier. Si le problème ne peut être formalisé à l'aide d'une fonction de coût, il demeure important de fournir au destinataire de l'étude une description détaillée de l'incertitude sur les quantités d'intérêt étudiées, ainsi que nous l'avons illustré dans le cas de l'estimation d'une hauteur de crue.

En conclusion, l'estimation bayésienne a l'avantage sur les approches classiques d'offrir une solution systématique pour prendre une décision optimale en présence d'incertitudes épistémiques. Le fait qu'elle nécessite la donnée d'une loi *a priori* peut apparaître comme une difficulté. Cependant, le fait de pouvoir intégrer par le biais d'une loi *a priori* une information experte, souvent disponible dans un contexte industriel, est plutôt un atout qu'un défaut. En l'absence d'une telle information, il est par ailleurs toujours possible d'utiliser un *a priori* de référence, tel qu'il a été développé J. Bernardo (2000).

Remerciements.

Nous remercions Pietro Bernardara (EDF R&D) d'avoir mis à notre disposition le jeu de données qui nous a permis d'illustrer notre propos. Ce travail a été partiellement financé par le Ministère de l'Economie, des Finances et de l'Industrie (DGCIS) dans le cadre du projet CSDL (Complex Systems Design Lab) du Pôle de Compétitivité System@tic Paris-Région.

Bibliographie

[1] E. de Rocquigny (2006). La maîtrise des incertitudes dans un contexte industriel, Journal de la Société Française de Statistique, 147(3):33–106.
[2] J. Bernier (2003). Décisions et comportement des décideurs face au risque hydrologique / Decisions and attitude of decision makers facing hydrological risk. Hydrological Sciences Journal, 43(3):301 – 316.
[3] E. Parent, J. Bernier. (2007) Le raisonnement Bayésien, Springer.
[4] J. Bernardo, (2000) Bayesian Theory, John Wiley.

Apprentissage et Modèles de Mélanges

Sélection de variables pour l'analyse discriminante, *Cathy Maugis, Gilles Celeux and Marie-Laure Martin-Magniette*

Une procédure de sélection de prédicteurs pour l'analyse discriminante gaussienne est proposée. Le problème est reformulé en un problème général de sélection de modèles. La modélisation suppose trois rôles possibles pour les prédicteurs : une variable peut être discriminante, linéairement dépendante d'une partie des variables discriminantes ou indépendante. Un critère de type BIC est utilisé pour résoudre le problème de sélection de modèles. L'identifiabilité de la collection de modèles et la consistance de la sélection de modèles sont établies. En pratique le rôle des variables est obtenu par un algorithme fondé sur deux algorithmes forward stepwise imbriqués. Des expériences sur données simulées et données réelles illustrent l'intérêt de cette procédure de sélection de variables.

Etude de l'association entre deux variables en présence de seuils de détection, *Hela Romdhani and Lamji Lakhal-Chaieb*

Nous nous intéressons à l'étude de l'association entre deux mesures sujettes à une censure à gauche fixe due à un seuil au dessous duquel les mesures ne peuvent être prises. Nous définissons (Romdhani, H., Lakhal-Chaieb, M. L. (2011 Submitted)) une version conditionnelle τ_b du τ de Kendall permettant de mesurer cette association. Nous développons un estimateur non paramétrique de τ_b et étudions ses propriétés asymptotiques. Sous un modèle de copule archimedienne, nous exprimons τ_b en fonction du paramètre de la copule. Nous en déduisons des estimateurs du paramètre de dépendance et du τ de Kendall global. Nous développons, ensuite, un test d'ajustement à une copule adapté à ce type de données. Nous étudions la performance des méthodes proposées avec des simulations et les illustrons par une application sur des données réelles sur le VIH.

Apprentissage non supervisé des structures des HMMs, *Rakia Jaziri, Mustapha Lebbah, Younés Bennani and Jean Hugues Chenot*

Nous proposons dans cet article une nouvelle approche hybride qui fait coopérer les cartes auto-organisatrices (SOM) et les chaînes de Markov cachées (HMMs) pour la modélisation de données structurées en séquences. La principale contribution de l'approche proposée consiste à extraire automatiquement la topologie d'un modèle de Markov caché sans aucune connaissance préalable du domaine d'application et tout en profitant de la topologie des séquences produites par la carte. Ce modèle (macro-HMM) est constitué d'un graphe de super-états, où chaque super-état représente un modèle de Markov caché secondaire (micro-HMM).L'approche proposée a été validée sur des données réelles de lettres manuscrites. Des résultats expérimentaux illustrent ses avantages.

Un algorithme EM normalisé pour les données directionnelles,

Wafia Parr Bouberima, Mohamed Nadif and Yamina Khemal Bencheikh

Dans ce papier nous traitons le problème de la classification de données directionnelles. Pour ce faire, nous utilisons un modèle des mélanges de lois exponentielles. Plaçant l'estimation des paramètres sous l'approche maximum de vraisemblance, nous utilisons un algorithme EM normalisé. Des expériences numériques réalisées sur des données réelles et simulées permettent d'évaluer les performances de l'algorithme proposé.

Sélection de variables pour l'analyse discriminante

Cathy Maugis[(1)] & Gilles Celeux[(2)] & Marie-Laure Martin-Magniette[(3,4)]

(1): Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France
(2) Inria Saclay Île-de-France, France
(3) UMR AgroParisTech/INRA MIA 518, Paris, France
(4) URGV UMR INRA 1165, UEVE, ERL CNRS 8196, Evry, France

Résumé:

Le problème de la sélection de variables pour la classification supervisée est un sujet bien étudié (voir Guyon and Ellisseeff,2003; Mary-Huard et al.,2007). On se focalise ici sur la sélection de variables pour des modèles génératifs gaussiens. Il existe des méthodes efficaces pour sélectionner les prédicteurs dans le cadre de l'analyse discriminante linéaire (LDA). Des procédures stepwise de sélection sont disponibles dans de nombreux logiciels de statistique (voir McLachlan, 1992, Section 12.3.3). Au contraire, il existe peu d'outils pour l'analyse discriminante quadratique (QDA)(Young and Odell, 1986), et à notre connaissance, aucune procédure de sélection de variables n'est disponible dans les logiciels standards. Cependant, on peut constater un regain d'intérêt pour cette thématique ces dernières années. Zhang et Wang (2008) ont proposé une procédure de sélection de variables pour QDA fondée sur un critère BIC et Murphy et al.(2010) ont adapté la procédure de Raftery et Dean (2006) dans le contexte de la classification supervisée.

Le but de notre travail est d'étendre la modélisation générale du rôle des variables proposée par Maugis et al.(2009b), conçue pour la classification non supervisée par mélanges gaussiens. Cette modélisation est le résultat d'améliorations successives proposées par Raftery and Dean (2006) and Maugis et al. (2009a,b). Nous désirons ainsi renforcer de façon significative l'intérêt des classifieurs gaussiens non linéaires (Bensmail et Celeux, 1996) qui sont actuellement limités par le grand nombre de paramètres à estimer. Les modèles et algorithmes de sélection de variables proposés permettent d'interpréter le rôle des variables mais aussi d'améliorer l'efficacité discriminante de méthodes telles que QDA.

Notre méthode de sélection des prédicteurs pour la classification générative gaussienne suppose qu'il existe trois types de variables : les variables discriminantes (S) utiles pour la classification, les variables redondantes (U) expliquées par un sous-ensemble (R) de variables discriminantes, et les variables indépendantes (W) qui n'apportent aucune information pour la classification. La sélection de variables est importante dans la procédure d'analyse discriminante pour obtenir un classifieur fiable et parcimonieux. Considérer le cadre de l'analyse discriminante gaussienne nous permet de reformuler le problème de sélection de variables en un problème général de sélection de modèles. Considérons les

données d'apprentissage composées de n vecteurs

$$(\underline{\mathbf{x}}, \underline{z}) = \{(\mathbf{x}_1, z_n), \ldots, (\mathbf{x}_n, z_n);\ \mathbf{x}_i \in \mathbb{R}^Q, z_i \in \{1, \ldots, K\}\},$$

où $\mathbf{x}_i$ est le prédicteur de dimension Q et z_i est le label du i^{eme} individu. La densité de l'échantillon d'apprentissage est modélisée par $\forall(\mathbf{x}, z) \in \mathbb{R}^Q \times \{1, \ldots, K\}$,

$$\begin{cases} f(\mathbf{x}|z = k, m, r, l, \mathbf{V}) & = & \Phi(\mathbf{x}^S|\mu_k, \Sigma_{k(m)})\, \Phi(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega_{(r)})\, \Phi(\mathbf{x}^W|\gamma, \tau_{(l)}) \\ \\ \quad (\mathbb{1}_{z=1}, \ldots, \mathbb{1}_{z=K}) & \sim & \text{Multinomial}(1; p_1, \ldots, p_K) \end{cases}$$

où $\Phi(.|\mu, \Sigma)$ dénote la densité d'une loi gaussienne multidimensionnelle de vecteur moyenne μ et de matrice de variance Σ. Les modèles considérés sont donc de la forme $(m, r, \ell, \mathbf{V})$ où $\mathbf{V} = (S, R, U, W)$ est la partition des variables, m désigne l'une des 14 formes disponibles en analyse discriminante par décomposition spectrale (Bensmail and Celeux, 1996), r est la forme de la matrice de variance dans la régression de U par rapport à R et ℓ est la forme de la matrice de variance dans la densité gaussienne modélisant la loi sur les variables indépendantes W. Le modèle sélectionné maximise alors le critère suivant :

$$\text{crit}(m, r, l, \mathbf{V}) = \text{BIC}_{\text{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) + \text{BIC}_{\text{reg}}(\underline{\mathbf{x}}^U|r, \underline{\mathbf{x}}^R) + \text{BIC}_{\text{indep}}(\underline{\mathbf{x}}^W|\ell).$$

Ce critère est la somme de trois critères BIC : le premier associé à l'analyse discriminante, le second à la régression linéaire et le troisième à la densité gaussienne indépendante.

D'un point de vue théorique, les conditions nécessaires et suffisantes pour assurer l'identifiabilité de la collection de modèles sont établies. Ces conditions assurent que la partie d'analyse discriminante, la régression linéaire et la densité gaussienne associée aux variables indépendantes se distinguent complètement. De plus, la consistance de la sélection de modèles est démontrée : la probabilité de sélectionner le vrai modèle (qui est supposé être dans la collection) en maximisant notre critère de sélection de modèles tend vers 1 lorsque la taille de l'échantillon tend vers l'infini. Ce résultat théorique est établi sous des hypothèses de bornitude des différents paramètres.

En pratique, le rôle des variables est obtenu par un algorithme fondé sur deux algorithmes forward stepwise imbriqués, l'un pour la sélection de variables en classification et l'autre pour celle en régression. Dans un premier temps, notre algorithme sépare les variables en deux catégories, les significatives et les non informatives pour l'analyse discriminante, via un algorithme forward stepwise. Dans une deuxième phase, les variables non informatives sont partitionnées en variables redondantes ou variables indépendantes selon si des régresseurs parmi les variables significatives peuvent les expliquer ou non par une régression linéaire. Finalement, on détermine les régresseurs parmi les variables significatives qui sont nécessaires pour expliquer les variables redondantes selon une régression multidimensionnelle.

Des expériences sur données simulées et sur données réelles illustrent l'intérêt de cette procédure de sélection de variables. En particulier, elles montrent l'intérêt d'une telle

modélisation du rôle des variables pour améliorer les performances de l'analyse discriminante quadratique dans un contexte de grande dimension.

Mots-Clés: variable significative, redondante ou indépendante, sélection de variables, modèles gaussiens pour la classification, régression linéaire, BIC.

Abstract:
A general methodology for selecting predictors for Gaussian generative classification models is proposed. The problem is regarded as a model selection problem. Three different roles for each possible predictor are considered: a variable can be a relevant classification predictor or not, and the irrelevant classification variables can be linearly dependent on a part of the relevant predictors or independent variables. This variable selection model was inspired by previous works on variable selection in model-based clustering (Maugis et al, 2009a,b; Raftery and Dean, 2006). A BIC-like model selection criterion is proposed. It is optimized through two embedded forward stepwise variable selection algorithms for classification and linear regression. The model identifiability and the consistency of the variable selection criterion are proved. Numerical experiments on simulated and real data sets illustrate the interest of this variable selection methodology. In particular, it is shown that this well ground variable selection model can be of great interest to improve the classification performance of the quadratic discriminant analysis in a high dimension context.

Keywords: Discriminant, redundant or independent variables, Variable selection, Gaussian classification models, Linear regression, BIC.

Bibliographie

[1] Bensmail, H. and Celeux, G. (1996) Regularized Gaussian Discriminant Analysis Through Eignenvalue Decomposition. *Journal of the American Statistical Association*, 91, 1743-1748.
[2] Guyon, I. and Ellisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning research*, 3, 1157-1182.
[3] Mary-Huard, T., Robin, S. and Daudin, J.-J. (2007) A penalized criterion for variable selection in classification. *Journal of Multivariate Analysis*, 98, 695-705.
[4] Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2009) Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65, 701-709.
[5] Maugis, C.,Celeux, G. and Martin-Magniette, M.-L. (2009) Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, 53, 3872-3882.

[6] McLachlan, G. (1992) *Discriminant Analysis and Statistical Pattern Analysis*. Wiley-Interscience, New York.
[7] Murphy, Brendan T., Raftery, Adrian E. and Dean, Nema (2010) Variable Selection and Updating in Model-Based Discriminant Analysis for High-Dimensional Data with Food Authenticity Applications. *Annals of Applied Statistics*, 4, 396-421.
[8] Raftery, A. and Dean, N. (2006) Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101, 168-178.
[9] Young, D. M. and Odell, P. L. (1986) Feature-subset selection for statistical classification problems involving unequal covariance matrices. *Communication in Statistics-Theory and Methods*, 15, 137-157.
[10] Zhang, Q. and Wang, H. (2008) A BIC Criterion for Gaussian Mixture Model Selection with Application in Discriminant Analysis. Technical report, Guanghua School of Management, Peking University.

On the association between left censored variables in presence of detection limits

Héla Romdhani & Lajmi Lakhal-Chaieb

Département de mathématiques et de statistique, Université Laval, 1045 Av. de la médecine, Québec, QC, Canada.

Investigating the association between different viral loads is the primary purpose of many HIV studies. Typically, a proportion of these loads measurements fall below detection limits due to the intensive use of highly active antiretroviral therapy. Such observations are said to be left-censored. Measuring the association between two variables subject to such censoring and studying its properties are the main subject of Romdhani, H., Lakhal-Chaieb, M.L. (2011 Submitted). Here we expose the methods proposed in this paper and their applications.

Let X and Y denote the two continuous random variables corresponding to the measurements with known lower detection limits L_X and L_Y, respectively. Due to left censoring, one may only observe n independent replications of $(\tilde{X}, \tilde{Y})$, where $\tilde{X} = \max(X, L_X)$ and $\tilde{Y} = \max(Y, L_Y)$. Some attempts have been made to provide nonparametric measures of association between X and Y under this setting. Typically, these are based on Kendall's tau, equal to $\tau = E[\psi_{12}]$, where $\psi_{12} = \text{sign}\{(X_1 - X_2)(Y_1 - Y_2)\}$, $\text{sign}(u) = -1$ if $u < 0$ and 1 if $u > 0$ and (X_1, Y_1) and (X_2, Y_2) are two independent replications of (X, Y). In the presence of fixed left-censoring, the relationship between X and Y is completely missing if $X \in [0, L_X]$ or $Y \in [0, L_Y]$ and therefore it is impossible to estimate τ nonparametically. However, one may estimate the conditional versions of Kendall's tau $\tau_b = E[\psi_{ij}|\xi_{ij}]$ where ξ_{ij} is the event $\{\max(X_i, X_j) > L_X, \max(Y_i, Y_j) > L_Y\}$. Note that ξ_{ij} denotes the orderability event. We adapt the methodology of [5] to the current setting and average the ψ_{ij}'s over the orderable pairs. The resulting estimator

$$\hat{\tau}_b = \frac{\sum_{i<j} I(\xi_{ij})\psi_{ij}}{\sum_{i<j} I(\xi_{ij})}, \tag{1}$$

is the empirical version of τ_b. We prove that $\sqrt{n}\{\hat{\tau}_b - \tau_b\}$ converges to $\mathcal{N}(0, \Gamma)$, with Γ consistently estimated by

$$\hat{\Gamma} = 2n^{-3} \sum_{k<l<m} \left\{\hat{Q}_{kl}\hat{Q}_{km} + \hat{Q}_{kl}\hat{Q}_{lm} + \hat{Q}_{lm}\hat{Q}_{km}\right\}$$

and $\hat{Q}_{ij}$ obtained by plugging in estimators for the unknown quantities in $Q_{ij} = \frac{2}{P(\xi_{12})} I(\xi_{ij})\{\psi_{ij} - \tau_b\}$. We show that, under an Archimedean copula model for (X, Y) with generator ϕ indexed by the dependence parameter α, τ_b is expressed as

$$\tau_b = \frac{\mathcal{C}_\alpha^2(p_x, p_y) - 4(p_x + p_y)\mathcal{C}_\alpha(p_x, p_y) + 4(p_x + p_y) - p_x^2 + p_y^2 + 1 + \mathcal{I}(p_x, p_y)}{1 - p_x^2 - p_y^2 + \mathcal{C}_\alpha^2(p_x, p_y)}, \tag{2}$$

where

$$\begin{aligned}
\mathcal{I}(a,b) &= 4\times\{h_1(a,b)+h_1(b,a)-h_2(a,b)+h_3(a,b)+h_3(b,a)\},\\
h_1(a,b) &= \int_0^{\phi_\alpha(a)}\frac{\phi_\alpha^{[-1]}\{t+\phi_\alpha(b)\}}{\phi'_\alpha\{\phi^{[-1]}(t)\}}dt,\\
h_2(a,b) &= \int_0^{\phi_\alpha(a)}\int_0^{\phi_\alpha(b)}\frac{1}{[\phi'_\alpha\{\phi_\alpha^{[-1]}(s+t)\}]^2}dsdt,\\
h_3(a,b) &= \int_0^{\phi_\alpha(a)}\frac{\phi_\alpha^{[-1]}(t)}{\phi'_\alpha[\phi_\alpha^{[-1]}\{t+\phi(b)\}]}dt,
\end{aligned}$$

and $p_x=\Pr(X<L_X)$ and $p_y=\Pr(Y<L_Y)$ are the censoring fractions.
Writing (2) as $\tau_b=g(\alpha,p_x,p_y)$, one may estimate α by the root of $\hat{\tau}_b=g(\alpha,\hat{p}_x,\hat{p}_y)$, where $\hat{\tau}_b$ is given by (1) and $\hat{p}_x$ and $\hat{p}_y$ are the empirical estimators of p_x and p_y. We prove that $\sqrt{n}(\hat{\alpha}-\alpha)=n^{-3/2}\sum_{i<j}R_{ij}+o_p(1)$, where

$$R_{ij}=\frac{2}{g_{100}(\alpha,p_x,p_y)}\{\frac{1}{P(\xi_{12})}I(\xi_{ij})(\psi_{ij}-\tau_b)-g_{010}(\alpha,p_x,p_y)(\frac{\gamma_i+\gamma_j}{2}-p_x)$$
$$-g_{001}(\alpha,p_x,p_y)(\frac{\beta_i+\beta_j}{2}-p_y)\},$$

$\gamma_k=I(X_k>L_X)$, $\beta_k=I(Y_k>L_Y)$ and $g_{lkm}(a,b,c)=\partial^{l+k+m}g(a,b,c)/\partial^l a\partial^k b\partial^m c$. The asymptotic normality and an estimator for the asymptotic variance of $\sqrt{n}(\hat{\alpha}-\alpha)$ follow.

With complete observations $\{(X_i,Y_i),i=1,\cdots,n\}$, Genest *et al.* [11] proposed a goodness-of-fit test based on the integral probability transformation $K(t)=\Pr\{\mathcal{C}(U,V)\le t\},0\le t\le 1$, where $U=F_X(X)$ and $V=F_Y(Y)$. We adapt the methodology of [11] to test the adequacy of an arbitrary Archimedean copula to a data set of observations with lower detection limits. Similarly to τ, K is not identifiable nonparametically under this setting because the missing information is not recoverable. Therefore, our procedure is based on a conditional version of K, somewhat analogous to τ_b

$$K_{\alpha,p_x,p_y}(t)=\Pr\{\pi(X,Y)\le t|X>L_X;Y>L_Y\}=\Pr\{\mathcal{C}_\alpha(U,V)\le t|U>p_x;V>p_y\}.$$

Its empirical version is

$$\hat{K}(t)=\frac{\sum_{i=1}^n I(X_i>L_X;Y_i>L_Y)I(\hat{T}_i\le t)}{\sum_{i=1}^n I(X_i>L_X;Y_i>L_Y)}. \tag{3}$$

We show that $\sqrt{n}\{\hat{K}(t) - K_{\alpha,p_x,p_y}(t)\}$ converges to a zero mean normal distribution for all $t \in [0,1]$. Furthermore, under an Archimedean copula model for (X,Y), we show that

$$K_{\alpha,p_x,p_y}(t) \times d = \begin{cases} 0 & \text{if } t \leq a \\ \frac{\phi_\alpha(t)-\phi_\alpha(b)-\phi_\alpha(c)}{\phi'_\alpha(t)} - t + a & \text{if } a \leq t \leq b \\ -\frac{\phi_\alpha(c)}{\phi'_\alpha(t)} - b + a & \text{if } b \leq t \leq c \\ t - \frac{\phi_\alpha(t)}{\phi'_\alpha(t)} - b - c + a & \text{if } t \geq c, \end{cases} \tag{4}$$

where $a = \mathcal{C}_\alpha(p_x, p_y)$, $b = \min(p_x, p_y)$, $c = \max(p_x, p_y)$ and $d = 1 + a - b - c$. A copula-based estimator of the conditional integral probability transformation is obtained from (4) by plugging in estimators for the parameters p_x, p_y and α.
Given a data set, one may conduct a goodness-of-fit test based on a comparison of $\hat{K}$ given by (3) and $K_{\hat{\alpha},\hat{p}_x,\hat{p}_y}$. Several metrics can be used to measure the distance between these curves. The limiting distribution of this test statistics is difficult to derive analytically. However, it is possible to use the parametric bootstrap to approximate the associated p-value.
The finite-sample performance of the proposed methods is evaluated by simulations and their use illustrated with a real data set on plasma and saliva viral loads.

Keywords: Archimedean copula, Goodness-of-fit test for copulas, Kendall's tau, Lower detection limits.

Etudier l'association entre différentes charges virales est le but de plusieurs travaux sur le VIH. Quand les sujets sont soumis à une thérapie antirétrovirale-hautement-active intensive, certaines mesures des charges virales ne peuvent être observées à cause de l'existence de seuils de détections. De telles observations sont dites censurées à gauche. Mesurer l'association entre deux variables sujettes à une telle censure et l'étude de ses propriétés sont le sujet de *Romdhani, H., Lakhal-Chaieb, M. L.(2011 Soumis)*. Nous exposons ici les méthodes proposées dans ce papier ainsi que leurs applications.

Soient X et Y deux mesures ayant des seuils de détection connus L_X et L_Y, respectivement. A cause de la censure à gauche, on ne peut observer que n copies indépendantes de $(\tilde{X}, \tilde{Y})$, où $\tilde{X} = \max(X, L_X)$ et $\tilde{Y} = \max(Y, L_Y)$. Certaines tentatives ont été faites pour fournir des mesures non paramétriques de l'association entre deux variables X et Y censurées à gauche. En général, elles sont basées sur le tau de Kendall qui est défini par $\tau = E[\psi_{12}]$, avec $\psi_{12} = \text{sign}\{(X_1 - X_2)(Y_1 - Y_2)\}$, $\text{sign}(u) = -1$ si $u < 0$ et 1 si $u > 0$ et (X_1, Y_1) et (X_2, Y_2) sont deux copies indépendantes de (X, Y). En présence de censure à gauche fixe, la relation entre X et Y est totalement inidentifiable si $X \in [0, L_X]$ ou $Y \in [0, L_Y]$. Il est donc impossible d'estimer τ d'une façon non paramétrique. Nous

pouvons, toutefois, estimer la version conditionnelle du tau de Kendall $\tau_b = E[\psi_{ij}|\xi_{ij}]$ où ξ_{ij} est l'événement $\{\max(X_i, X_j) > L_X, \max(Y_i, Y_j) > L_Y\}$. Notons que ξ_{ij} dénote l'événement d'ordorabilité. Nous adaptons la méthodologie de [5] à notre type de données en conditionnant par rapport à cet événement d'ordorabilité. L'estimateur obtenu

$$\hat{\tau}_b = \frac{\sum_{i<j} I(\xi_{ij})\psi_{ij}}{\sum_{i<j} I(\xi_{ij})}, \tag{5}$$

est la version empirique de τ_b. Nous prouvons que $\sqrt{n}\{\hat{\tau}_b - \tau_b\}$ converge vers la distribution $\mathcal{N}(0, \Gamma)$, avec

$$\hat{\Gamma} = 2n^{-3} \sum_{k<l<m} \left\{\hat{Q}_{kl}\hat{Q}_{km} + \hat{Q}_{kl}\hat{Q}_{lm} + \hat{Q}_{lm}\hat{Q}_{km}\right\}$$

un estimateur consistant de Γ, où $\hat{Q}_{ij}$ est obtenu en remplaçant dans $Q_{ij} = \frac{2}{P(\xi_{12})} I(\xi_{ij})\{\psi_{ij} - \tau_b\}$ les termes inconnus par leurs estimateurs. Nous montrons que, sous un modèle de copule archimedienne pour (X, Y) de générateur ϕ indexé par le paramètre de dépendance α, τ_b s'écrit comme

$$\tau_b = \frac{\mathcal{C}_\alpha^2(p_x, p_y) - 4(p_x + p_y)\mathcal{C}_\alpha(p_x, p_y) + 4(p_x + p_y) - p_x^2 + p_y^2 + 1 + \mathcal{I}(p_x, p_y)}{1 - p_x^2 - p_y^2 + \mathcal{C}_\alpha^2(p_x, p_y)}, \tag{6}$$

où

$$\begin{aligned}
\mathcal{I}(a, b) &= 4 \times \{h_1(a, b) + h_1(b, a) - h_2(a, b) + h_3(a, b) + h_3(b, a)\}, \\
h_1(a, b) &= \int_0^{\phi_\alpha(a)} \frac{\phi_\alpha^{[-1]}\{t + \phi_\alpha(b)\}}{\phi'_\alpha\{\phi^{[-1]}(t)\}} dt, \\
h_2(a, b) &= \int_0^{\phi_\alpha(a)} \int_0^{\phi_\alpha(b)} \frac{1}{[\phi'_\alpha\{\phi_\alpha^{[-1]}(s + t)\}]^2} ds dt, \\
h_3(a, b) &= \int_0^{\phi_\alpha(a)} \frac{\phi_\alpha^{[-1]}(t)}{\phi'_\alpha[\phi_\alpha^{[-1]}\{t + \phi(b)\}]} dt,
\end{aligned}$$

et $p_x = \Pr(X < L_X)$ et $p_y = \Pr(Y < L_Y)$ sont les fractions de censure.
En écrivant (6) comme $\tau_b = g(\alpha, p_x, p_y)$, on peut estimer α par la racine de l'équation $\hat{\tau}_b = g(\alpha, \hat{p}_x, \hat{p}_y)$, où $\hat{\tau}_b$ est donné par (5) et $\hat{p}_x$ et $\hat{p}_y$ les estimateurs empirique de p_x et p_y. Nous prouvons aussi que $\sqrt{n}(\hat{\alpha} - \alpha) = n^{-3/2} \sum_{i<j} R_{ij} + o_p(1)$, où

$$R_{ij} = \frac{2}{g_{100}(\alpha, p_x, p_y)} \{\frac{1}{P(\xi_{12})} I(\xi_{ij})(\psi_{ij} - \tau_b) - g_{010}(\alpha, p_x, p_y)(\frac{\gamma_i + \gamma_j}{2} - p_x)$$

$$- g_{001}(\alpha, p_x, p_y)(\frac{\beta_i + \beta_j}{2} - p_y)\},$$

$\gamma_k = I(X_k > L_X)$, $\beta_k = I(Y_k > L_Y)$ et $g_{lkm}(a,b,c) = \partial^{l+k+m} g(a,b,c)/\partial^l a \partial^k b \partial^m c$. La normalité asymptotique et un estimateur de la variance asymptotique de $\sqrt{n}(\hat{\alpha} - \alpha)$ s'en suivent.

Dans le cas de données complètes $\{(X_i, Y_i), i = 1, \cdots, n\}$, Genest *et al.* [11] ont proposé un test d'ajustement basé sur la fonction $K(t) = \Pr\{\mathcal{C}(U,V) \leq t\}, 0 \leq t \leq 1$, où $U = F_X(X)$ et $V = F_Y(Y)$. Nous adaptons la méthodologie of [11] pour tester l'adéquation d'unw copule Archimedienne quelconque à un ensemble de données en présence de seuils de détection. D'une façon similaire à τ, K n'est pas identifiable d'une façon non paramétrique pour ce type de données. Par conséquent, notre procédure est basée sur une version conditionnelle de K, en quelque sorte analogue à τ_b

$$K_{\alpha,p_x,p_y}(t) = \Pr\{\pi(X,Y) \leq t | X > L_X; Y > L_Y\} = \Pr\{\mathcal{C}_\alpha(U,V) \leq t | U > p_x; V > p_y\}.$$

Sa version empirique est

$$\hat{K}(t) = \frac{\sum_{i=1}^n I(X_i > L_X; Y_i > L_Y) I(\hat{T}_i \leq t)}{\sum_{i=1}^n I(X_i > L_X; Y_i > L_Y)}. \tag{7}$$

Nous montrons que $\sqrt{n}\{\hat{K}(t) - K_{\alpha,p_x,p_y}(t)\}$ converge vers la distribution normale centrée pour tout $t \in [0,1]$. Nous prouvons, en outre, que sous un modèle de copule archimedienne pour (X,Y), on a

$$K_{\alpha,p_x,p_y}(t) \times d = \begin{cases} 0 & \text{si } t \leq a \\ \frac{\phi_\alpha(t) - \phi_\alpha(b) - \phi_\alpha(c)}{\phi'_\alpha(t)} - t + a & \text{si } a \leq t \leq b \\ -\frac{\phi_\alpha(c)}{\phi'_\alpha(t)} - b + a & \text{si } b \leq t \leq c \\ t - \frac{\phi_\alpha(t)}{\phi'_\alpha(t)} - b - c + a & \text{si } t \geq c, \end{cases} \tag{8}$$

où $a = \mathcal{C}_\alpha(p_x, p_y)$, $b = \min(p_x, p_y)$, $c = \max(p_x, p_y)$ et $d = 1 + a - b - c$. Un estimateur de la fonction $K(t)$, basé sur la copule, est obtenu en remplaçant dans (8), les paramètres p_x, p_y et α par leurs estimateurs.

Nous pouvons alors effectuer, sur un jeux de données, un test d'ajustement basé sur une comparaison de $\hat{K}$ donné par (7) et $K_{\hat{\alpha},\hat{p}_x,\hat{p}_y}$. Pour mesurer la distance entre ces deux courbes, plusieurs métriques peuvent être utilisées. La distribution asymptotique de ces statistiques de test sont particulièrement difficiles à obtenir analytiquement. Il est, par contre, possible d'utiliser le bootstrap paramétrique pour approximer la p-value associée au test.

La performance, à taille finie, des méthodes proposées est évaluée à l'aide de simulations et leur utilisation illustrée par une application sur des données réelles sur le VIH.

Mots-clés: Copule Archimedienne, Tau de Kendall, Test d'ajustement de copules, Seuil de détection.

Bibliographie

[1] Romdhani, H., Lakhal-Chaieb, M. L. (2011 Submitted) On the association between variables with lower detection limits. *Statistics in Medecine* [2] Barroso, P. F., Schechter, M., Gupta, P., Melo, M. F., Vieira, M., Murta, F. C., Souza, Y. and Harrison, L. H.(2000). Effect of antiretroviral therapy on HIV shedding in semen. *Ann. Intern. Med.* **133**, 280-284.
[3] Nie, L., Chu, H. and Kororstyshevskiy, V. R. (2008). Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits. *The Canadian Journal of Statistics* **36**, 427–442.
[4] Chu, H., Nie, L. and Zhu, M. (2008). On estimation of bivariate biomarkers with known detection limits. *Environmetrics* **19**, 301–317.
[5] Gibbons J.D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference.* 4th ed., Marcel Dekker Inc: New York.
[6] Oakes, D. (2008). On consistency of Kendalls tau under censoring. *Biometrika* **95**, 997–1001.
[7] Lyles, R.H., Williams, J.K. and Chuachoowong, R. (2001). Correlating two viral load assays with known detection limits. *Biometrics* **57**, 1238–1244.
[8] Chu, H., Moulton, L. H., Mack, W. J., Passaro, D. J., Barraso, P. F. and Munoz, A. (2005). Correlating two continuous variables subject to detection limits in the context of mixture distributions. *Appli. Statist.* **54**, 831–845.
[9] Wang, A. (2007). The Analysis of Bivariate Truncated Data Using the Clayton Copula Model. *The International Journal of Biostatistics*, **Vol 3**, Iss. 1, Article 8.
[10] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Pub. Inst. Statist. Univ. Paris*, **8**, 229–231.
[11] Manatunga, A.K. and Oakes D. (1996). A meaure of association for bivariate frailty distributions. *Journal of Multivariate Analysis* **56**, 60–74.
[12] Genest, C. Quessy, J.-F. and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian journal of statistics*, **33**, 337–366.
[13] Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Inst. Henri Poincar—'e A*, **44**, 1096–1127.
[14] Barbe, P., Genest, C., Ghoudi, K. and Rémillard, B. (1996). On Kendall's process. *Journal of Multivariate Analysis* **58**: 197–229.

Apprentissage non supervisé des structures des HMMs

Rakia Jaziri, Mustapha Lebbah, Younès Bennani et Jean-Hugues Chenot

Laboratoire d'Informatique de Paris-Nord, Université Paris 13
99, av. J-B Clément, F-93430 Villetaneuse
{*prénom.nom*}*@lipn.univ-paris13.fr*
Institut National de l'Audiovisuel
4, av. de l'Europe 94 366 Bry-sur-Marne
{*rjaziri,jhchenot*}*@ina.fr*

Nous proposons dans cet article une nouvelle approche hybride qui fait coopérer les cartes auto-organisatrices (SOM) et les chaînes de Markov cachées (HMMs) pour la modélisation de données structurées en séquences. La principale contribution de l'approche proposée consiste à extraire automatiquement la topologie d'un modèle de Markov caché sans aucune connaissance préalable du domaine d'application et tout en profitant de la topologie des séquences produites par la carte auto-organisatrice. Ce modèle (macro-HMM) est constitué d'un graphe de super-états, où chaque super-état représente un modèle de Markov caché secondaire (micro-HMM). L'approche proposée a été validée sur des données réelles de lettres manuscrites. Des résultats expérimentaux illustrent ses avantages.
Mots clés: modèle de Markov caché, modèles de mélanges, cartes auto-organisatrices.

1 Introduction

Les modèles de Markov cachés, Baum et al. (1970), figurent parmi les meilleures approches adaptées aux traitements des séquences, étant donnée leur capacité à traiter des séquences de longueurs variables, et leur pouvoir à modéliser la dynamique d'un phénomène décrit par des suites d'événements. Ce modèle est défini par une structure composée d'états et de transitions et par un ensemble de distributions de probabilité sur les transitions. Ils sont largement utilisés dans de nombreux domaines de la reconnaissance des formes, la modélisation des séquences biologiques, voir Durbin et al (1998), la reconnaissance vocale et la reconnaissance optique des caractères pour n'en nommer que quelques-uns. De toute évidence, il y a beaucoup de structures possibles de modèles de Markov qui peuvent être construites en fonction des besoins des applications. L'une des importantes restrictions des HMM est leur incapacité à extraire automatiquement l'architecture des modèles sans aucune connaissance préalable du domaine d'application. Il est donc important d'avoir des algorithmes capables de déduire, à partir d'un ensemble de données de séquences, non seulement la probabilité de distribution, mais aussi la structure topologique du modèle. Malheureusement, cette tâche est très difficile et il n'existe que des solutions partielles. Par exemple dans Bouchaffra, et Tan (2006) et Bouchaffra (2008), les auteurs proposent un paradigme original, appelé *topological HMM*, qui manipule les noeuds du graphe associé au HMM et ses transitions dans un espace Euclidien. D'autres approches ont été proposées, consistant à combiner les HMMs et les cartes auto-organisatrices (Self-Organizing Map,

Kohonen (2001)), pour obtenir un modèle hybride SOM-HMM dans lequel chaque cellule de la carte représente un HMM, voir Ferles et Stafylopatis (2008). Cependant, le processus d'organisation n'est pas intégré explicitement dans l'approche HMM. Notre papier est organis comme suit : La section 2 présente la nouvelle approche proposée, qui combine les points forts des HMM et SOM. La section 3 décrit le dispositif expérimental et les évaluations. Enfin, la section 4 présente une conclusion et des perspectives d'évolution.

2 Apprentissage non supervisé des structures des HMM

Dans notre modèle, nous considérons une chaîne de Markov $\lambda = (A, B, \Pi, C)$ où $C = \{\mathbf{c}_1, ..., \mathbf{c}_K\}$ est l'ensemble des états, $\mathbf{A}$ est la matrice des probabilités de transition entre les états, $\mathbf{B}$ est l'ensemble des paramètres de la distribution qui est connue comme la probabilité d'émission, et $\Pi = \{\pi_1, ..., \pi_K\}$ est l'ensemble des probabilités initiales. On note par $\mathbf{X}$ un vecteur de séquences de taille N, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, ..., \mathbf{x}_N\}$, où $\mathbf{x}_n$ est un élément de la séquence. L'idée que nous proposons consiste à utiliser le pouvoir d'auto-organisation des cartes pour guider la découverte de la structure des HMMs. Afin d'introduire notre modèle, nous définissons un HMM sous forme d'une grille de cellules C, qui a une topologie discrète définie par un graphe non orienté. Ce graphe sera modélisé par une carte topologique composée de K cellules corespondant chacune à un état du micro-HMM. Pour chaque paire d'état (c, r) sur la carte, la distance $\delta(c, r)$ est définie comme le plus court chemin reliant les deux états c et r. Ainsi toutes les probabilités permettant de définir le HMM seront estimées à partir des paramètres de la carte topologique.

Nous considérons que la distribution initiale des états dépend de la cardinalité de chaque cellule (nombre d'éléments affectés à chaque cellule). Cette probabilité est estimée par : $\pi_i = \frac{card(c_i)}{|X|}, 1 \leq i \leq K$; où $card\,(c_i)$ est le nombre d'éléments affectés à la cellule c_i et $|X|$ représente la taille de la séquence. La génération d'une variable observable $\mathbf{x}_i$ dans l'état c_i à un instant donné du temps est conditionnée par les états voisins c_j au même instant. Cette proximité est quantifiée en utilisant la fonction de voisinage $\mathcal{K}$, $\mathcal{K}\left(\delta\left(c_i, c_j\right)\right) = e^{\frac{\delta\left(c_i, c_j\right)}{T}}, T_{min} \leq T \leq T_{max}$ avec $\delta\left(c_i, c_j\right)$ la distance de Manhattan sur la carte entre deux cellules, T est un paramètre de tempèrature pour contrôler le voisinage

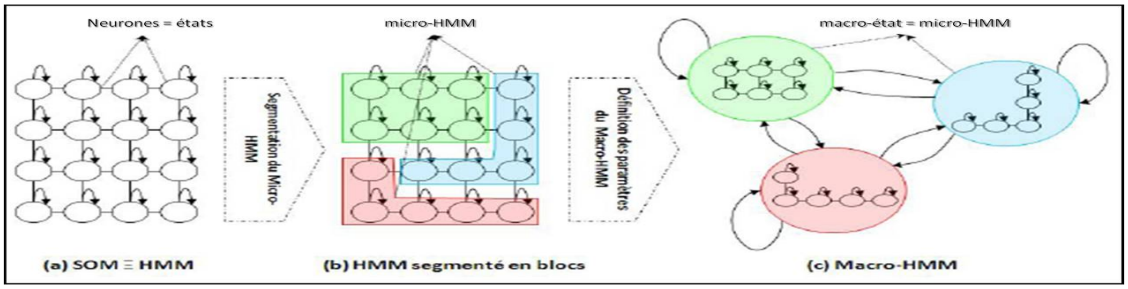


Figure 1: Approche de modélisation des cartes topologiques et du modèle markovien.

dans le temps. Ainsi la probabilité de transition entre c_i et c_j sera définie par :

$$a_{i,j} = P\left(c_i/c_j\right) = \frac{\mathcal{K}\left(\delta\left(c_i, c_j\right)\right)}{\sum_{i=1}^{K} \mathcal{K}\left(\delta\left(c_i, c_j\right)\right)}, 1 \leq i, j \leq K$$

La probabilité d'émission à un état c_i est définie par une gaussienne d'écart type σ_i et du centre $\mathbf{w}_{c_i}$. $P\left(\mathbf{x}_i/c_j\right) = \frac{1}{\sqrt{2\pi}\sigma_{c_j}} \exp \frac{-||\mathbf{x}_i - \mathbf{w}_{c_j}||^2}{2\sigma_{c_j}^2}, 1 \leq i \leq N$ et $1 \leq j \leq K$

La carte topologique est ensuite segmentée en S clusters par l'application d'une Classification Ascendante Hiérarchique (CAH) sur les prototypes $\mathbf{w}_{c_i}$. Chaque cluster modélisera un super-état (macro-HMM) S_k d'un modèle appelé macro-HMM dont les paramètres seront aussi calculés à partir de la carte topologique et du résultat de la segmentation ainsi défini. En d'autres termes, chaque état du macro-HMM λ^* est composé de plusieurs états initials (cellules de la carte). La probabilité initiale d'un état du macro-HMM est déterminée par la somme des probabilités initiales π des états composant: $\pi_k = \sum_{j \in S_k} \pi_j$.

Afin de définir les probabilités de transition du macro-HMM, chaque super-état est associé à un micro-état c^* représentatif du cluster. Cet état c^* est le plus proche en terme de distance euclidienne des centres $\mathbf{w}_c$ des états formant le cluster. La probabilité de transition, $a_{i,j}$, est calculée aussi à l'aide de la fonction $\mathcal{K}(\delta(c_i^*, c_j^*))$ modélisant le voisinage entre les deux représentants. Elle est définie comme suit :

$$a_{i,j} = P\left(c_i^*/c_j^*\right) = \frac{\mathcal{K}^T\left(\delta\left(c_i^*, c_j^*\right)\right)}{\sum_{c_i} \mathcal{K}^T\left(\delta\left(c_i^*, c_j^*\right)\right)}, 1 \leq i, j \leq S$$

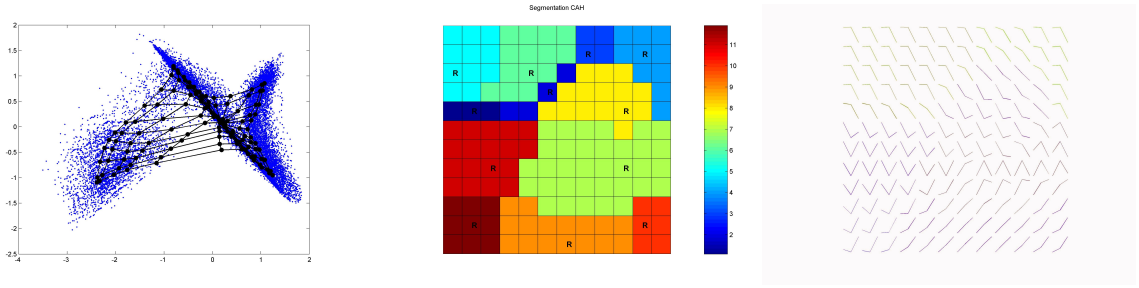
Chaque super-état S_k est un modèle de mélange à $|S_k|$ composantes. La probabilité d'émission est calculée par la formule suivante : $P\left(\mathbf{x}/S_k\right) = \sum_{c_j \in S_k} P\left(c_j/c_k^*\right) P\left(\mathbf{x}/c_j\right)$ Après cette première estimation, une seconde phase d'apprentissage est lancée à partir de la base intiale. Cette phase permet un raffinement des probabilités de transition a_{ij}, et donc une amélioration des estimations initiales.

3 Expérimentation

Dans cette section, nous présentons une application de notre approche sur une base de données réelle issue du répertoire UCI, Asuncion et Newman (2007). Les données se composent de 2858 séquences. Elles ont été capturées à l'aide d'une tablette WACOM, où les 3 dimensions, x, y, et la force de pointe du stylo, ont été conservées. Chaque caractère est une trajectoire de vitesse de pointe du stylo. Il s'agit d'un contenu sous forme de matrice, avec 3 lignes (x, y, z)et N colonnes, où N est la longueur de la séquence.

Dans la première étape de notre expérimentation, nous avons modélisé notre approche macro-HMM en apprenant les lettres séparément. La figure 2.a représente la projection des données avec la carte apprise pour la lettre 'a'. Ces projections fournissent une visualisation topographique des données séquentielles. Les points en bleu représentent les éléments de la séquence originale. La figure 2.b présente la segmentation de la carte en différents clusters ou chaque cluster présente un macro-état composé de plusieurs micro-état. Nous notons R la position de l'état sélectionné c^*. La figure 2.c montre les profils

des cellules du micro-HMM. Chaque état représente les 3 variables: vitesse x, vitesse y et la force de pointe du stylo. Nous remarquons que les prototypes associés à chaque cluster sont similaires et correspondant au découpage de la carte.



(a) (b) (c)

Figure 2: carte 12×12. Projection dans le plan des composantes, x et y, des données de la lettre 'a'.(a) Les points en bleu représentent les éléments de la séquence,(b) Segmentation de la carte et selection des représentants 'R'(c) Prototypes associés aux clusters.

Une des caractéristiques qui distingue notre modèle macro-HMM des autres HMM, est la préservation du voisinage produite par la carte. Notons aussi que le nombre d'états et la structure du modèle final sont déterminés automatiquement. La figure 3 montre l'architecture du HMM déduite automatiquement à partir de la carte topologique.

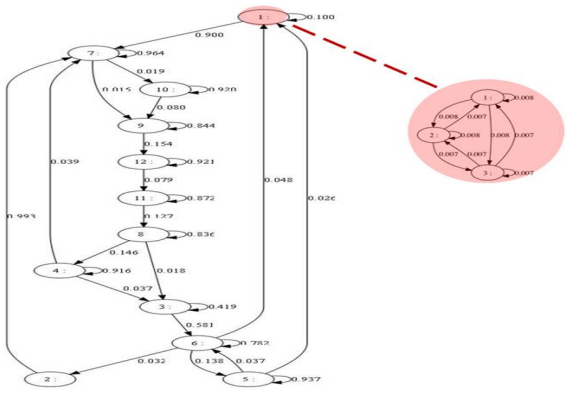
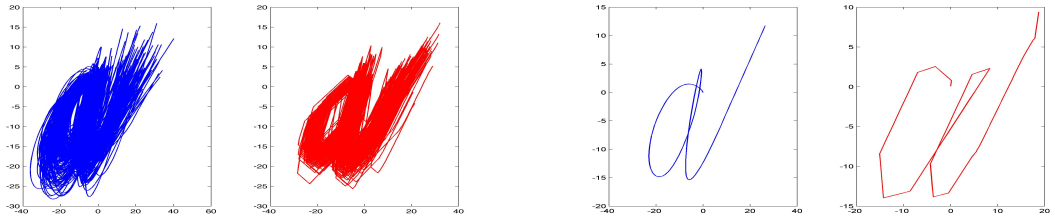


Figure 3: L'architecture du macro-HMM correspondant à la lettre 'a'.

Afin de mettre en évidence le bon déroulement de la phase d'apprentissage, nous avons calculé pour cet exemple le chemin de Viterbi le plus probable. Le caractère coloré en bleu est l'échantillon original et celui coloré en rouge est l'échantillon reproduit. Le graphique

à gauche de la figure 4.a présente les différents exemples d'apprentissage de la lettre 'a' et le graphique à droite de la même figure indique les formes générées par notre modéle (macro-HMM). La figure 4.b indique la même chose pour un seul exemple.



(a) (b)

Figure 4: (a) Reconstruction de la lettre avec l'ensemble des échantillons. (b) Reconstruction de la lettre 'a' avec un échantillon.

Dans la seconde étape de notre expérimentation, nous avons testé notre modèle comme classifieur en utilisant la technique de validation croisée. Nous avons subdivisé la base en cinq sous-ensembles de données. A chaque itération, nous utilisons quatre sous ensembles pour l'apprentissage, et le reste pour la phase de test. Les étiquettes générées ont été comparées aux étiquettes réelles pour chaque base de test. Dans la table I, nous observons que l'utilisation des cartes topologiques améliore les performances du modèle et réduit la variance des résultats. Nous avons comparé nos résultats avec celles d'un HMM ergodique, voir Bishop, C. M. (2006), possédant le même nombre d'état. Les résultats obtenus montrent que notre modèle fournit en générale une structure plus performante.

Table 1: Performances de la validation croisée sous forme d'intervalles. Taux de bonne classification

Lettres/modèle	HMM	Macro-HMM
Lettre 'a'	[96.38 - 99.53]	[98.48 - 100]
Lettre 'b'	[97.25 - 99.81]	[98.48 - 100]
Lettre 'h'	[80.77 - 89.46]	[96.04 - 99,40]
Lettre 'm'	[64.46 - 75.71]	[98.48 - 100]
Lettre 'n'	[57.71 - 69.54]	[98.48 - 100]

4 Conclusion

Les travaux présentés dans ce papier portent sur l'apprentissage des structures des HMM sans connaissances a priori du domaine d'application. Il s'agit d'une approche qui détermine une structure optimisée des HMMs à partir de la topologie des séquences. Cette topologie est détectée automatiquement par apprentissage non supervisé d'une carte topologique. La structure des HMM identifiée est adaptée à la structure et à la qualité des données.

L'approche proposée a été évaluée sur un jeu de données de lettres manuscrites. Les résultats obtenus sont encourageants et montrent l'efficacité de l'apprentissage automatique de la structure. La topologie extraite décrit non seulement la dynamique des séquences de données grâce aux transitions entre les macro-états, mais aussi les différentes intra-structures des séquences. Nous envisagons maintenant d'appliquer l'approche proposée pour le traitement des données audio-visuelles.

Abstract: We propose in this paper a novel approach which makes self-organizing maps (SOM) and the Hidden Markov Models (HMMs) cooperate. The main contribution for the proposed approach is to automatically extract the structure of a hidden Markov model without any prior knowledge of the application domain. This model can be represented as a graph of macro-states, where each state represents a micro model. Experimental results illustrate the advantages of the proposed approach, compared to a fixed structure approach.

Bibliographie

[1] Asuncion, A., Newman, D. (2007): UCI machine learning repository.

[2] Baum, L.E. Petrie, T. Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains, The Annals of Mathematical Statistics, 41, pp. 164-171, .

[3] Bishop, C. M. (2006) Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA).

[4] Bouchaffra, D. and Tan, J. (2006) Structural hidden markov models : An application to handwritten numeral recognition. Intell. Data Anal.10(1) :67-79.

[5] Bouchaffra, D. (2008) Embedding hmm's-based models in a euclidean space : The topological hidden markov models. In ICPR08, pages 1-4,.

[6] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G.(1998). Biological sequence analysis. Cambridge University Press.

[7] Ferles, C. and Stafylopatis, A. (2008) "Sequence clustering with the Self-Organizing Hidden Markov Model Map", BioInformatics and BioEngineering journal, pp.1-7.

[8] Kohonen, T. (2001) Self-organizing Maps. Springer Berlin.

[9] Rabiner, L. R. (1989) A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257286.

UN ALGORITHME EM NORMALISÉ POUR LES DONNÉES DIRECTIONNELLES

Wafia Parr Bouberima [1,2], Mohamed Nadif [1], Yamina Khemal Bencheikh [2]
wboub@yahoo.fr, mohamed.nadif@parisdescartes.fr, bencheikh_00@yahoo.fr

1. *LIPADE, UFR Maths-Info, Université Paris Descartes, 45, rue des Saints Pères 75006 Paris, France.*
2. *Laboratoire LMFN, Département de Mathématiques, Faculté des Sciences, Université Ferhat Abbas, Setif. Algérie.*

Résumé : Dans ce papier nous traitons le problème de la classification de données directionnelles. Pour ce faire, nous utilisons un modèle de mélange de lois exponentielles. Plaçant l'estimation des paramètres sous l'approche maximum de vraisemblance, nous utilisons un algorithme EM normalisé. Des expériences numériques réalisées sur des données réelles et simulées permettent d'évaluer les performances de l'algorithme proposé.

Abstract: In this paper we focus on clustering of directional data. To this end, we use a mixture model of exponential distributions. We use a normalized EM algorithm to estimate the parameters of the model. Numerical experiments on real and synthetic data allow to evaluate the performances of this algorithm.

Mots clés : Modèles de Mélange, EM Normalisé

1 Introduction

Nous supposons que les données initiales se présentent sous forme d'un tableau rectangulaire de taille $n \times d$. La classification automatique d'un ensemble de n objets décrits par un ensemble de d attributs peut être réalisée en utilisant l'approche modèle de mélange. Celle-ci est devenue ces vingt dernières années une approche classique, voir par exemple l'ouvrage de McLachlan et Peel (2000).

Dans cette approche les données sont supposées provenir d'un échantillon de mélange de composants, qui sont modélisés par une distribution de probabilité. Les algorithmes souvent proposés sont basés généralement sur la maximisation d'une vraisemblance des données observées et sont de type EM (Dempster et al. (1977)), ou encore sur la maximisation d'une vraisemblance classifiante et dans ce cas une version classifiante de EM (CEM) est utilisée (Celeux and Govaert (1992)).

Les données directionnelles apparaissent souvent dans de nombreuses disciplines scientifiques (voir par exemple, Mardia and Jupp (2000)) telles que l'analyse textuelle et les données biopuces. Pour le traitement de ce type de données, l'algorithme k-means sphérique proposé par Dhillon et Modha (2001) et destiné au traitement des données

textuelles, utilise le cosinus de similarité entre des vecteurs normés, pour obtenir une partition structurée sur l'ensemble des objets. Celle-ci notée $\mathbf{z}$ est représentée par une matrice de classification binaire $\mathbf{z} = (z_{ik})$ de taille $(n \times g)$. Le critère à optimiser par cet algorithme dans ce cas s'écrit :

$$W(\mathbf{z}, \mu) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \mathbf{x}_i^T \mu_k,$$

où $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T$ et $(\mu_k)_{k=1,...,g}$ centres des classes appartiennent à l'hypersphère S^{d-1} de rayon 1. Notons que ce critère est associé à un modèle de mélange de distributions de Von Mises-Fisher sous contraintes (Banerjee et al. (2005)). En supposant que l'ensemble des objets est un échantillon i.i.d qui provient d'un mélange de g composants, la vraisemblance des données s'écrit

$$f(\mathbf{x}; \theta) = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k \varphi_k(\mathbf{x}_i; \alpha_k),$$

où les π_k sont les proportions du mélange vérifiant : $\sum_{k=1}^{g} \pi_k = 1, \pi_k \geq 0$, $\alpha_k = (\mu_k, \xi_k)$, μ_k et ξ_k désignent respectivement la moyenne directionnelle et le paramètre de concentration des vecteurs autour de cette moyenne.

$\varphi_k(\mathbf{x}_i; \alpha_k) = c_d(\xi_k) \exp(\xi_k \mathbf{x}_i^T \mu_k)$ avec $c_d(\xi) = \frac{\xi^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\xi)}$ où $I_{\frac{d}{2}}(\xi)$ est la fonction de Bessel modifiée du 1^{er} type d'ordre $\frac{d}{2}$, définie par : $I_d(\xi) = \frac{1}{2\pi} \int_0^{2\pi} \cos d\theta e^{\xi \cos\theta} d\theta$.

Des algorithmes de type EM ont été développés, cependant les résultats ont été souvent décevants lorsque la dimension est très grande. Pour cette raison, Phuong et Vinh (2008) ont proposé un modèle de mélanges parcimonieux défini sur une sphère de rayon quelconque fixé au départ. Cette démarche a permis en particulier de montrer que la qualité des partitions obtenues dépend foncièrement du rayon de la sphère. Dans notre travail, nous allons utiliser ce modèle en s'affranchissant de la connaissance à priori de ce paramètre.

2 Modèle de mélange

Les données utilisées sont normalisées (au sens de la norme L_2) de manière à appartenir à une sphère S_ρ de rayon ρ

$$S_\rho = \{\mathbf{x} | \, \|\mathbf{x}\|^2 = \rho, \mathbf{x} \in R^d\}.$$

Autrement dit, chaque vecteur $\mathbf{x}_i$ est remplacé par $\sqrt{\rho} \frac{\mathbf{x}}{||\mathbf{x}||}$. L'ensemble des objets étant supposé constitué un échantillon i.i.d et les objets sont supposés provenir d'un mélange de g composants de distributions exponentielles de paramètres $(\mu_k), k = 1, \ldots, g$, de la forme suivante :

$$\varphi_k(\mathbf{x}|\theta) = \gamma_\rho \sum_{k=1}^{g} \pi_k e^{\|\mathbf{x} - \mu_k\|^2}.$$

Les μ_k sont les moyennes directionnelles associées aux classes et vérifient : $\|\mu_k\|^2 = \rho, k = 1, \ldots, g$, avec γ_ρ une fonction constante de normalisation de la forme :

$$\gamma_\rho = \frac{1}{\int_{\mathbf{x} \in S_\rho} e^{\|\mathbf{x}-\mu_k\|^2} d\mathbf{x}}.$$

Dans (Phuong et Vinh (2008)), les auteurs en supposant que le rayon de la sphère est constant (choisi par l'utilisateur) ont proposé un algorithme EM normalisé. Cependant, en variant ce paramètre, les auteurs ont noté à partir de plusieurs essais que la qualité des partitions dépendait fortement de ce paramètre. Pour cette raison, dans la suite nous allons considérer que le rayon est inconnu et sera estimé au cours des itérations, ainsi $\theta = (\pi, \mu, \rho)$.

3 Algorithme EM normalisé

L'algorithme itératif EM remplace la maximisation de la log-vraisemblance par son espérance de données conditionnellement aux données observées et l'estimation courante $\theta^{(c)}$. Celle-ci s'écrit

$$Q(\theta, \theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log(\pi_k \varphi_k(\mathbf{x}_i; \alpha_k))$$

où $s_{ik}^{(c)} = P(z_{ik} = 1 | \mathbf{x}_i, \theta^{(c)}) = \frac{\pi_k^{(c)} \varphi_k(\mathbf{x}_i; \alpha_k^{(c)})}{\sum_{k'=1}^{g} \pi_{k'}^{(c)} \varphi_{k'}(\mathbf{x}_i; \alpha_{k'}^{(c)})}$ est la probabilité a posteriori qu'un $\mathbf{x}_i$ provienne du composant k.

L'objectif des estimations des paramètres du modèle décrit précédemment est réalisé grâce à EM qui se décompose en l'itération de deux étapes : Estimation et Maximisation. Dans le contexte modèle de mélange, l'étape *Estimation* se réduit aux calculs des probabilités a posteriori $s_{ik}^{(c)}$; dans l'étape *Maximisation*, on recherche le paramètre θ maximisant

$$Q(\theta, \theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log\left(\pi_k \gamma_\rho e^{\|\mathbf{x}_i - \mu_k\|^2}\right)$$

ou encore,

$$Q(\theta, \theta^{(c)}) = n \ln \gamma_\rho - 2ng\rho + \sum_{i,k} s_{ik}^{(c)} (\ln \pi_k + 2\mathbf{x}_i^T \mu_k).$$

On peut montrer que les paramètres $(\pi_k, \mu_k, \rho), k = 1, \ldots, g$ sont définis de la manière suivante :

- $\pi_k^{(c)} = \frac{\sum_i s_{ik}^{(c)}}{n}$
- $\mu_k^{(c)} = \sqrt{\rho} \frac{\sum_i s_{ik}^{(c)} \mathbf{x}_i}{\left\| \sum_{i=1}^{n} s_{ik}^{(c)} \mathbf{x}_i \right\|}$
- Le rayon de la sphère $\rho^{(c)}$ est solution de l'équation : $\rho[g + \frac{4\rho \exp(-4\rho^2)}{1 - \exp(-4\rho^2)}] = \frac{2}{n} \sum_{i,k} s_{ik}^{(c)} \mathbf{x}_i^T \mu_k^{(c)}$.

A la convergence de l'algorithme, les classes sont déduites par le principe du maximum à posteriori.

4 Conclusion

Dans ce travail, nous traitons le problème de la classification de données directionnelles sous l'approche modèle de mélange. Nous utilisons un modèle parcimonieux. Les paramètres du modèle sont les proportions des classes, les centres des classes et le rayon de la sphère dans laquelle sont projetés les objets et les centres. L'algorithme EM normalisé utilisé est stable et capable de surmonter la grande dimension.

5 Bibliographie

Banerjee, A., Dhillon, I. S., Ghosh, J. et Sra, S. (2005), "Clustering on the unit hypersphere using von Mises-Fisher distributions", Journal of Machine Learning Research, 6:1345-1382.
Celeux, G. et Govaert, G. (1992), "A Classification EM Algorithm for Clustering and Two Stochastic versions". Computational Statistics and Data Analysis, 14, 315-332.
Dempster, A. P., Laird N. M. et Rubin, D. B. (1977), Maximum likelihood for incomplet data via the EM algorithme". Journal of the royal statistical society, 39(B):1-38.
Dhillon, I. S. et Modha, D. S. (2001), "Concept decompositions for large sparse text data using clustering", Machine Learning, 42(1):143-175.
McLachlan, G.J. et Peel, D. (2000), *Finite mixture models*, Wiley, New York.
Mardia, K.V. et Jupp, P.E. (2000), *Directional Statistics*, Wiley, Chichester.
Phuong, N.M. et Vinh, N.X. (2008), "Normalized EM algorithm for tumor clustering using gene expression data", In Proc. the 8th IEEE International conference on Bioinformatics and BioEngineering, Athens 8-10th Oct 2008.

Statistique Mathématique

Bornes de sparsité en suites individuelles dans un cadre de régression linéaire séquentielle, *Sebastien Gerchinovitz*

On s'intéresse au problème de la régression linéaire séquentielle en grande dimension pour des suites déterministes arbitraires. Dans ce cadre, on prouve des bornes de regret qui sont un équivalent déterministe des inégalités oracle de sparsité introduites au cours de la dernière décennie dans un cadre stochastique. Notre algorithme séquentiel SeqSEW procède par mélange exponentiel et troncature dépendante des données. Dans un second temps, on applique une version totalement automatique de cet algorithme au modèle de régression avec design aléatoire. Dans ce cadre, les bornes obtenues pour des suites individuelles impliquent des inégalités oracle de sparsité exactes qui sont comparables à celles de Dalalyan et Tsybakov (2008, 2010), mais qui répondent à deux questions soulevées par ces auteurs. En particulier, nos bornes de risque sont adaptatives en la variance inconnue du bruit (à un facteur logarithmique près) si ce dernier est Gaussien.

Testing for conditional symmetry in absolutely regular and possibly nonstationary dynamical models, *Joseph Ngatchou-Wandji and Michel Harel*

We propose a symmetry test for the errors distribution in a class of heteroscedastic models. The observations as well as the errors are not necessarily stationnary, but are required to be absolutely regular. The convergence of the residual-based empirical distribution function is established, as well as some other functional limit theorems. The null cumulative distribution function of the test statistic is approximated. A simulation experiment shows that the test performs well on the example tested.

Une propriété d'indépendance caractérisant le produit des lois de Kummer et gamma, *Efoevi Koudou and Pierre Vallois*

Koudou et Vallois (2010) ont caractérisé le produit de la loi de Kummer et de la loi gamma par une propriété d'indépendance, sous l'hypothèse de l'existence de densités deux fois différentiables et strictement positives. Nous démontrons la même caractérisation sous l'hypothèse plus faible de l'intégrabilité locale des logarithmes des densités. La méthodologie de la preuve est inspirée par l'article de Wesolowski (2002) concernant la propriété de Matsumoto-Yor caractérisant le produit des lois gaussienne inverse généralisée et gamma.

Distribution of the determinant of a sample correlation matrix IN and applications, *Thu Pham-Gia*

La matrice de corrélation en statistique multidimensionnelle est fréquemment utilisée dans les applications. Cependant sa distribution exacte est seulement connue quand la matrice de

corrélation de la population est l'Identité, sauf dans le cas d'une distribution bivariée. Aussi, la distribution de son déterminant est inconnue. Dans cette communication nous étudions cette dernière distribution avec les fonctions spéciales de Meijer G, qui sont à présent calculables. Nous l'appliquons ensuite au concept de dépendance d'un système. Pour le cas général quand la matrice de corrélation de la population n'est pas l'Identité une étude par simulation procure quelques résultats.

Echantillonnage conditionnel, *Virgile Caron and Michel Broniatowski*

Nous présentons un schéma d'approximation de la densité d'un vecteur aléatoire (X1,...,Xk) conditionnée à la valeur observée de (X1 + ... + Xn)/n lorsque k/n -> 1 ; on traite également le cas où le conditionnement est de la forme (f(X1)+...+f(Xn))/n. Les conditionnements considérés sont supposés "normaux", c.a.d. dans la gamme des valeurs impliquées par la loi du log-itéré. La qualité de l'approximation est étudiée et un algorithme explicite fournit la valeur maximale de k, en fonction de n, pour laquelle une précision relative fixée est atteinte. L'approximation proposée permet la simulation d'échantillons (Y1,...,Yk) sous le conditionnement induit par l'observation (f(X1)+...+f(Xn))/n. Nous utilisons cet échantillonnage conditionnel pour la tabulation de tests relatifs à un paramètre d'intérêt sous un paramètre de nuisance, lorsque le conditionnement est définie par la valeur prise par une statistique exhaustive pour ce paramètre de nuisance. Ces méthodes sont aussi applicables pour l'amélioration d'estimateurs dans le cadre du théorème de Rao-Blackwell.

Bornes de sparsité en suites individuelles dans un cadre de régression linéaire séquentielle

Sébastien Gerchinovitz

DMA, Ecole Normale Supérieure[1]*, 45 rue d'Ulm, 75005 Paris.*
sebastien.gerchinovitz@ens.fr

Résumé On s'intéresse au problème de la régression linéaire séquentielle en grande dimension pour des suites déterministes arbitraires. Dans ce cadre, on prouve des bornes de regret qui sont un équivalent déterministe des inégalités oracle de sparsité introduites au cours de la dernière décennie dans un cadre stochastique. Notre algorithme séquentiel SeqSEW procède par mélange exponentiel et troncature dépendante des données. Dans un second temps, on applique une version totalement automatique de cet algorithme au modèle de régression avec design aléatoire. Dans ce cadre, les bornes obtenues pour des suites individuelles impliquent des inégalités oracle de sparsité exactes qui sont comparables à celles de Dalalyan et Tsybakov [3, 4], mais qui répondent à deux questions soulevées par ces auteurs. En particulier, nos bornes de risque sont adaptatives en la variance inconnue du bruit (à un facteur logarithmique près) si ce dernier est Gaussien.

Mots-clés : apprentissage séquentiel, suites individuelles, sparsité, bornes de regret.

Abstract We consider the problem of online linear regression on arbitrary deterministic sequences when the ambient dimension d can be much larger than the number of time rounds T. We introduce the notion of sparsity regret bound, which is a deterministic online counterpart of recent risk bounds derived in the stochastic setting under a sparsity scenario. We prove such regret bounds for an online-learning algorithm called SeqSEW and based on exponential weighting and data-driven truncation. In a second part we apply a parameter-free version of this algorithm on i.i.d. data and derive risk bounds of the same flavor as in [3, 4] but which solve two questions left open therein. In particular our risk bounds are adaptive (up to a logarithmic factor) to the unknown variance of the noise if the latter is Gaussian.

Index terms : online learning, individual sequences, sparsity, regret bounds.

1. Cette recherche a été conduite au sein du projet INRIA CLASSIC hébergé par l'École Normale Supérieure et le CNRS.

1 Introduction

Au cours de la dernière décennie, le phénomène de sparsité a fait l'objet de nombreux travaux dans le cadre stochastique. Parmi les outils introduits à cet effet, la notion d'*inégalité oracle de sparsité* – ou *sparsity oracle inequality* en anglais – joue un rôle fondamental. En régression linéaire, de telles bornes impliquent que la tâche consistant à prédire presque aussi bien qu'un vecteur inconnu de grande dimension est statistiquement faisable pourvu que ce vecteur ait peu de coordonnées non nulles.

Dans cette communication, on introduit un équivalent séquentiel déterministe de la notion d'inégalité oracle de sparsité. Les bornes déterministes obtenues seront, dans un second temps seulement, appliquées au cas particulier de suites i.i.d. ; cela permettra de répondre à deux questions soulevées par Dalalyan et Tsybakov [4].

1.1 Modèle séquentiel déterministe

Le cadre que nous considérons est celui de la prévision de suites déterministes arbitraires (ou *suites individuelles*). Un statisticien doit prédire de façon séquentielle, à chaque tour $t = 1, 2, \ldots$, la valeur $y_t \in \mathbb{R}$ d'une suite inconnue d'observations en fonction d'une valeur d'entrée $x_t \in \mathcal{X}$ et de prédicteurs de base $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$, à partir desquels il formule sa propre prévision $\widehat{y}_t \in \mathbb{R}$. La qualité des prévisions est évaluée avec la perte carrée. L'objectif du statisticien est de prédire presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{d} u_j \varphi_j$, où $\boldsymbol{u} \in \mathbb{R}^d$, i.e., de satisfaire, uniformément sur toutes les suites individuelles $(x_t, y_t)_{1 \leqslant t \leqslant T}$, une borne de regret de la forme

$$\sum_{t=1}^{T} \bigl(y_t - \widehat{y}_t\bigr)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\bigr)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\} ,$$

pour un terme de regret $\Delta_{T,d}(\boldsymbol{u})$ aussi petit que possible et, en particulier, sous-linéaire en T.

1.2 Scénario de sparsité

Dans le cadre décrit ci-dessus, une variante de l'algorithme de prévision séquentielle Ridge étudié par Azoury et Warmuth [1] et Vovk [8] assure un regret d'ordre au plus $d \ln T$. Quand la dimension ambiante d est bien plus grande que le nombre de tours de prévision T, cette dernière borne de regret est malheureusement bien supérieure à T et est donc en quelque sorte triviale. Puisque la borne $d \ln T$ est optimale en un certain sens (cf. [8, Théorème 2]), des hypothèses supplémentaires sont nécessaires pour garantir des performances théoriques intéressantes.

Une hypothèse naturelle, qui a déjà été maintes fois étudiée dans le cadre stochastique, est qu'il existe une combinaison linéaire sparse $\boldsymbol{u}^*$ (i.e., avec $s \ll T/(\ln T)$ coordonnées non

nulles) dont la perte cumulée est petite. Si le statisticien connaissait à l'avance le support $\{j : u_j^* \neq 0\}$ de $\boldsymbol{u}^*$, il pourrait appliquer le même algorithme de prévision séquentielle que précédemment mais seulement au sous-espace vectoriel de dimension s donné par $\big\{\boldsymbol{u} \in \mathbb{R}^d : \forall j \notin J(\boldsymbol{u}^*), u_j = 0\big\}$. Le regret de cet "oracle" serait alors au plus de l'ordre de $s \ln T$ et donc sous-linéaire en T. Sous ce "scénario de sparsité", un regret sous-linéaire semble donc possible, même si, bien sûr, la borne de regret $s \ln T$ peut seulement être utilisée comme une borne idéale de référence (puisque le support de $\boldsymbol{u}^*$ est inconnu).

Dans cette communication et sa version papier étendue [6], on montre qu'une borne de regret proportionnelle à s est atteignable (à un facteur logarithmique près). On prouve ainsi en Théorèmes 1 et 2 des bornes de regret de la forme

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \big(\|\boldsymbol{u}\|_0 + 1\big)\, g_{T,d}\big(\|\boldsymbol{u}\|_1, \|\boldsymbol{\varphi}\|_\infty\big) \right\}, \qquad (1)$$

où $\|\boldsymbol{u}\|_0$ dénote le nombre de coordonnées non nulles de $\boldsymbol{u}$ et où g est croissante mais croît au plus logarithmiquement en T, d, $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^{d} |u_j|$, et $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$. Nous appellerons *bornes de regret de sparsité* les bornes de regret de la forme précédente.

1.3 Travaux connexes dans les cadres stochastique et déterministe

La borne de regret (1) peut être vue comme un équivalent séquentiel déterministe des *inégalités oracle de sparsité* introduites par Bunea et al. dans le cadre stochastique au cours de la dernière décennie (cf., par ex, [2]). Le lecteur est invité à se reporter à la version étendue [6] pour de plus amples références sur cette abondante littérature.

Mentionnons néanmoins que, récemment, depuis les travaux de Dalalyan et Tsybakov [3], des inégalités oracle de sparsité avec constante 1 ont été prouvées sans presque aucune hypothèse sur le dictionnaire $(\varphi_j)_j$, et pour des méthodes procédant par pondération exponentielle qui peuvent être approchées numériquement à un coup algorithmique raisonable pour de grandes valeurs de la dimension ambiante d.

Quant au cadre séquentiel déterministe, à notre connaissance, les Théorèmes 1 et 2 (cf. aussi [6, Theorem 1]) fournissent les premiers exemples de borne de regret de sparsité au sens de (1). De récents travaux [7, 5] en optimisation convexe séquentielle ont certes abordé la question de la sparsité, mais sous un tout autre angle. Dans le cas de la régularisation ℓ^1 sous la perte carrée, ces travaux proposent des algorithmes qui prédisent comme une combinaison linéaire sparse $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{\varphi}(x_t)$ des prévisions de base (i.e., $\|\widehat{\boldsymbol{u}}_t\|_0$ est petit), alors que de telles garanties ne semblent pas pouvoir être montrées pour notre algorithme SeqSEW. En revanche, ils prouvent des bornes sur le regret ℓ^1-régularisé de la forme

$$\sum_{t=1}^{T}\Big((y_t - \widehat{\boldsymbol{u}}_t\cdot\boldsymbol{x}_t)^2 + \lambda\,\|\widehat{\boldsymbol{u}}_t\|_1\Big) \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^{T}\Big((y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + \lambda\,\|\boldsymbol{u}\|_1\Big) + \widetilde{\Delta}_{T,d}(\boldsymbol{u}) \right\},$$

pour un terme de regret $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ qui croît trop rapidement (comme une puissance et non logarithmiquement) en la dimension ambiante d, en la norme $\|\boldsymbol{u}\|_1$ ou en T. Ces algorithmes sont donc sous-optimaux dans le cadre qui nous intéresse ici (prévision sur des boules ℓ^0 de petit diamètre), ce qui n'est pas le cas des méthodes satisfaisant une borne de regret de la forme (1) comme, par exemple, notre algorithme SeqSEW.

2 Bornes de sparsité en suites individuelles

Pour simplifier l'analyse, on suppose d'abord que, au début du jeu, le statisticien a accès au nombre T de tours de prévision, à une borne B_y sur toutes les observations $y_1, \dots, y_T$ et à une borne B_Φ sur la trace de la matrice de Gram empirique, i.e.,

$$y_1, \dots, y_T \in [-B_y, B_y] \qquad \text{et} \qquad \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi \ .$$

La première version de notre algorithme est définie en Figure 1. Nous l'appelons *SeqSEW* puisqu'il s'agit d'une variante adaptée aux suites individuelles de l'algorithme Sparse Exponential Weighting introduit dans le cadre stochastique par Dalalyan et Tsybakov [3].

Paramètres : seuil $B > 0$, température inverse $\eta > 0$ et résolution $\tau > 0$ à laquelle on associe le prior π_τ sur $\mathbb{R}^d$ défini par

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^{d} \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1 + |u_j|/\tau\big)^4} \ .$$

Initialisation : $p_1 \triangleq \pi_\tau$.

A chaque tour de prévision $t \geqslant 1$,

1. Recevoir la donnée x_t et prédire $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_B \, p_t(\mathrm{d}\boldsymbol{u})$,
   où $[x]_B \triangleq \max\big\{-B, \min\{B, x\}\big\}$;
2. Recevoir l'observation y_t et calculer la probabilité a posteriori p_{t+1} sur $\mathbb{R}^d$ via l'expression (W_{t+1} est une constante de renormalisation)

$$p_{t+1}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{1}{W_{t+1}} \exp\left(-\eta \sum_{s=1}^{t} \Big(y_s - \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u}).$$

FIGURE 1 – Définition de l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$.

Au moyen d'une analyse PAC-Bayésienne (déterministe), on montre que cet algorithme satisfait la borne de regret de sparsité suivante.

Théorème 1 *Supposons que, pour des constantes connues $B_y > 0$ et $B_\Phi > 0$, les $(x_1, y_1), \dots, (x_T, y_T)$ soient tels que $y_1, \dots, y_T \in [-B_y, B_y]$ et $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$.*

Alors, l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *calibré avec $B = B_y$, $\eta = \dfrac{1}{8B_y^2}$ et $\tau = \sqrt{\dfrac{16B_y^2}{B_\Phi}}$ satisfait*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 32\, B_y^2\, \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\,\|\boldsymbol{u}\|_1}{4\, B_y\, \|\boldsymbol{u}\|_0}\right) \right\} + 16 B_y^2 \ .$$

Si le statisticien n'a pas accès à une borne a priori B_y sur les observations, il peut s'adapter séquentiellement à cette borne inconnue en tronquant les prévisions $\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)$ de façon dépendante des données. L'algorithme plus sophistiqué SeqSEW_τ^* prédit ainsi selon

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t} p_t(\mathrm{d}\boldsymbol{u}) \ , \qquad \text{où} \quad B_t \triangleq \inf\left(\left\{ \sqrt{2^k}; k \in \mathbb{Z} \right\} \cap \Big[\max_{1\leqslant s\leqslant t-1} |y_s|, +\infty \Big) \right) \ ,$$

et où la probabilité a posteriori p_t sur $\mathbb{R}^d$ est définie comme précédemment mais en remplaçant la température η par $\eta_t \triangleq 1/(8B_t^2)$ et, pour chaque indice s de la somme, le seuil B par B_s. Une analyse PAC-Bayésienne plus approfondie conduit à la borne suivante.

Théorème 2 *Pour tout $\tau > 0$, l'algorithme* SeqSEW_τ^* *précédent satisfait la borne*

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 64 \left(\max_{1\leqslant t\leqslant T} y_t^2\right) \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\, \tau}\right) \right\} \\ + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 32 \max_{1\leqslant t\leqslant T} y_t^2 \ .\end{aligned}$$

Au vu du dernier théorème, le choix du paramètre τ requiert encore la connaissance a priori d'une borne B_Φ sur $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t)$. Cela peut être évité au moyen d'un "doubling-trick" sur cette double somme, qui donne encore lieu à une borne de regret de sparsité.

3 Adaptativité en la variance pour des données i.i.d.

Dans cette section, on applique l'algorithme SeqSEW au modèle de régression avec design aléatoire (le cas du design fixe peut être traité similairement, cf. [6]). Le statisticien a accès à T copies indépendentes $(X_1, Y_1), \dots, (X_T, Y_T)$ de $(X, Y) \in \mathcal{X}\times\mathbb{R}$ de loi inconnue. On suppose que $\mathbb{E}[Y^2] < \infty$; l'objectif du statisticien est d'estimer la fonction de régression $f : \mathcal{X} \to \mathbb{R}$ définie par $f(x) \triangleq \mathbb{E}[Y|X = x]$ pour tout $x \in \mathcal{X}$. On pose aussi $\|h\|_{L^2} \triangleq \big(\mathbb{E}[h(X)^2]\big)^{1/2}$ pour toute fonction mesurable $h : \mathcal{X} \to \mathbb{R}$ telle que $\mathbb{E}[h(X)^2] < \infty$.

L'échantillon $(X_t, Y_t)_{t=1}^T$ est traité de façon séquentielle, en appliquant l'algorithme SeqSEW_τ^* de la date 1 à la date T avec $\tau = 1/\sqrt{dT}$. L'estimateur $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ retenu est défini par

$$\widehat{f}_T(x) \triangleq \frac{1}{T} \sum_{t=1}^{T} \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\big]_{B_t} \, p_t(\mathrm{d}\boldsymbol{u}) \ .$$

Nous prouvons dans [6] un théorème valable sous de faibles hypothèses sur la loi de Y. Mentionnons seulement le cas sous-gaussien ci-dessous.

Théorème 3 *Il existe une constante absolue $C > 0$ telle que, si $\|f\|_\infty < +\infty$ et, pour une constante $\sigma^2 > 0$ inconnue, $\mathbb{E}\Big[e^{\lambda(Y_1 - f(X_1))} \;\Big|\; X_1\Big] \leqslant e^{\lambda^2\sigma^2/2}$ p.s., alors, pour tout $T \geqslant 2$,*

$$\mathbb{E}\left[\left\| f - \widehat{f}_T \right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + 2C\left(\|f\|_\infty^2 + \sigma^2 \ln T\right) \frac{\|\boldsymbol{u}\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\ \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\}$$
$$+ \frac{1}{dT} \sum_{j=1}^{d} \|\varphi_j\|_{L^2}^2 + \frac{C}{T}\left(\|f\|_\infty^2 + \sigma^2 \ln T\right) \ .$$

Cette borne est comparable à la Proposition 1 prouvée par Dalalyan et Tsybakov [4]. Elle vaut néanmoins sur $\mathbb{R}^d$ tout entier au lieu de boules ℓ^1 de rayons finis, ce qui résout une question laissée ouverte dans [4, Section 4.2]. Par ailleurs, notre algorithme ne requiert pas la connaissance a priori du facteur de variance $\sigma^2 > 0$ du bruit, ce qui résout une seconde question soulevée dans [4, Section 5.1, Remark 6].

Références

[1] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3) :211–246, 2001.

[2] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4) :1674–1697, 2007.

[3] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2) :39–61, 2008.

[4] A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2010. Submitted.

[5] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.

[6] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. Technical report, 2011. See `http://arxiv.org/abs/1101.1057`.

[7] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10 :777–801, 2009.

[8] V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69 :213–248, 2001.

Testing for conditional symmetry in absolutely regular and possibly nonstationary dynamical models

Joseph Ngatchou-Wandji[a] & Michel Harel[b,*]

[a] *EHESP, Rennes and Université Henri Poincaré, Nancy, France*
[b] *IIUFM du Limousin, Limoges Cedex 87036, France*
[*] *IMT (UMR CNRS 5219), Université Paul Sabatier, Toulouse Cedex 31062, France.*

Let $U_i = \{(Y_i, X_i),\ i \in \mathbb{N}\}$ be an absolutely regular and nonnecessarily stationary process. The condional mean and variance of Y_i given Z_i=(Y_{i-1},Y_{i-2},...,Y_{i-s},X_i,X_{i-1},...,X_{i-q}) are denoted $m(Z_i;\theta)$ and $\sigma^2(Z_i;\rho)$, where m and σ are specified functions, ρ and θ unknown parameters. Denote $U^\top$ the transpose of a vector or matrix U. Consider the standardized form

$$\varepsilon_i(\psi) = \frac{Y_i - m(Z_i;\rho)}{\sigma(Z_i;\theta)},\ \psi = (\rho^\top, \theta^\top)^\top \in \Theta \times \widetilde{\Theta} \subset \mathbb{R}^l \times \mathbb{R}^p \tag{1}$$

of Y_i with cumulative distribution function F_i.

Let $\mathcal{S} = \{G$ continue cumulative distribution function: $G(x) = 1 - G(-x),\ x \in \mathbb{R}\}$. Let F be the limit of the sequence F_i. Our goal is to generalize the procedure proposed in Ngatchou-Wandji (2009) to testing the null hypothesis $\mathcal{H}_0 :\ F \in \mathcal{S}$ verses $\mathcal{H}_1 :\ F \notin \mathcal{S}$.

Such a work can be motivated by the facts that : in adaptive estimation, the hypothesis of symmetry of the error is needed for obtaining efficient bounds; in ARCH models, the effect of heteroscedasticity can depend on the symmetry of the error; many time series encountered in practice are not stationary.

Define the following random functions :

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\varepsilon_i(\psi_n)\le t)},\ t \in \mathbb{R} \tag{2}$$

$$S_n(t) = n^{1/2}\int_{\mathbf{Supp(F)}} \sin(tx)\omega(t)d\widehat{F}_n(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \sin[t\varepsilon_i(\psi_n)]\omega(t),\ t \in \mathbb{R},$$

where **Supp(F)** is the support of the distribution associated with F, $\psi_n = (\rho_n^\top, \theta_n^\top)^\top$ is a consistent estimator of the true parameter $\psi_0 = (\rho_0^\top, \theta_0^\top)^\top$, $\omega(t)$ a positive continuous function such that $\sup_{t\in\mathbf{Supp(F)}} |t\omega(t)| < \infty$.

From the study of the asymptotic properties of the process $S_n(t)$, many test statistics can be derived. However, here we only study the Cramér-von Mises type test based on the statistic

$$\mathcal{T}_n = \int_{\mathbf{Supp(F)}} S_n^2(t)d\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} S_n^2[\varepsilon_i(\psi_n)].$$

We show that for all $i \in \mathbb{N}$, $\varepsilon_i(\psi_n)$ and $\varepsilon_i(\psi_0)$ are asymptotic in probability, the empirical process $\widehat{F}_n(t)$ of the $\varepsilon_i(\psi_n)$'s converges uniformly in probability to $F(t)$, and under $\mathcal{H}_0$ the process $S_n(t)$ converges in distribution to a mean-zero Gaussian process $S(t)$ with covariance kernel Σ_0 depending of the parameters of the model and having a very general form. It results that under $\mathcal{H}_0$, $\mathcal{T}_n$ converges in distribution to a weighted sum of independent chi-square random variables with one degree of freedom, the weights in that sum being the eigenvalues of the integral operator ∇_{Σ_0} defined for all function g satisfying $\int_{\mathbb{R}} g^2(s)dF(s) < \infty$ by :

$$\nabla_{\Sigma_0} g(t) = \int_{\mathbb{R}} \Sigma_0(s,t)g(s)dF(s). \tag{3}$$

Proceeding as in Ngatchou-Wandji (2009), these eigenvalues can be estimated and the convergence of the estimators established. The p-values of the test can next be approximated by using, for example, Imhof (1961)'s results.

Résumé - Nous proposons un test de symétrie de la loi du bruit dans une classe de modèles hétéroscédastiques. Le cadre de cette étude est celui où, aussi bien les observations que les erreurs, ne sont plus nécessairement stationnaires, mais absolument régulières. Nous établissons la convergence du processus empirique des résidus et d'autres théorèmes limites fonctionnelles. Ces résultats permettent l'approximation des quantiles de la loi asymptotique de la statistique de test sous l'hypothèse nulle. Des simulations numériques montrent que notre test est performant sur l'exemple considéré.

References

[1] Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. Biometrika, 48, 419-426.

[2] Ngatchou-Wandji, J. (2009). Testing symmetry of the error distribution in nonlinear heteroscedastic models. Communications in Statistics - Theory and Methods, 38:1465-1485.

An independence property for the product of Kummer and gamma laws

A. E. Koudou & P. Vallois

Institut Elie Cartan, Laboratoire de Mathématiques, B.P. 239,
F-54506 Vandoeuvre-lès-Nancy cedex,
Efoevi.Koudou@iecn.u-nancy.fr, Pierre.Vallois@iecn.u-nancy.fr

Résumé : Koudou et Vallois (2010) ont caractérisé le produit de la loi de Kummer et de la loi gamma par une propriété d'indépendance, sous l'hypothèse de l'existence de densités deux fois différentiables et strictement positives. Nous démontrons la même caractérisation sous l'hypothèse plus faible de l'intégrabilité locale des logarithmes des densités. La méthodologie de la preuve est inspirée par l'article de Wesolowski (2002) concernant la propriété de Matsumoto-Yor caractérisant le produit des lois gaussienne inverse généralisée et gamma.

Abstract: Koudou and Vallois (2010) proved an independence property characterizing the product of Kummer and gamma distributions, assuming the existence of twice differentiable and strictly positive densities. This paper gives the same characterization under the weaker assumption of local integrability of the logarithms of the densities. Our proof is inspired by the Wesolowski (2002) work about the Matsumoto-Yor property characterizing a product of generalized inverse Gaussian and gamma laws.

Keywords: Gamma distribution; generalized inverse Gaussian distribution; Matsumoto-Yor property; Kummer distribution.

1 Introduction

Consider the gamma distribution $\gamma(\lambda, c)(dx) = \frac{c^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-cx} \mathbf{1}_{(0,\infty)}(x)dx, \quad (\lambda, c > 0)$, the beta distribution $\mathrm{Beta}(a, b)(dx) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\mathbf{1}_{\{0<x<1\}}dx, \quad (a, b > 0)$ and (see for instance Ng and Kotz, 1995, or Nagar and Gupta, 2002 and references therein) the *Kummer distribution of type 2* :

$$K^{(2)}(a, b, c) := \alpha(a, b, c)x^{a-1}(1 + x)^{-a-b}e^{-cx}\mathbf{1}_{(0,\infty)}(x)dx, \ a, c > 0, \ b \in \mathbb{R} \tag{1.1}$$

where $\alpha(a, b, c)$ is a normalizing constant.

Recall that the GIG distribution with parameters $\mu \in \mathbb{R}$, $a, b > 0$ is the probability measure $\mathrm{GIG}(\mu, a, b)$ with density proportional to $x^{\mu-1}e^{-\frac{1}{2}(a^2x^{-1}+b^2x)}\mathbf{1}_{(0,\infty)}(x)$.

Let f be a decreasing function from $(0, \infty)$ onto $(0, \infty)$. Given two independent, positive random variables X and Y, consider $U = f(X + Y)$ and $V = f(X) - f(X + Y)$. For $f(x) = \log\left(\frac{1+x}{x}\right)$, Koudou and Vallois (2010) proved that U and V are independent if and only if there exist constants a, b, c such that X follows the Kummer distribution of type

2, $K^{(2)}(a,b,c)$ and Y the gamma distribution $\gamma(b,c)$. The authors made the assumption that the r.v.'s have twice differentiable density functions. The case $f(x) = 1/x$ is named in the literature *the Matsumoto-Yor property*. According to this property, (partially proved by Matsumoto and Yor (2001) and completely shown by Letac and Wesolowki (2000)), for $f(x) = 1/x$, the r.v.'s U and V are independent if and only if there exist constants $a, b, \mu > 0$ such that $X \sim \text{GIG}(-\mu, a, b)$ and $Y \sim \gamma(\mu, b^2/2)$. The only requirement was that the r.v.'s be non-Dirac. The independence of the r.v.'s X, Y and of U, V is characterized by a functional equation involving their densities and the function f (see Equation (3.10) below. For $f(x) = 1/x$, Wesolowski (2002) solved that functional equation and, by the way, rediscovered the Matsumoto-Yor property under the assumption that the logarithms of the densities are locally integrable. We realized that their method can be applied to the Kummer-gamma case.

We state the result in Section 1 and prove it in Section 2.

2 The result

Define on $(0,\infty)$, $f(x) = \frac{1}{e^x - 1}$ and its inverse $g(x) = f^{-1}(x) = \ln\left(1 + \frac{1}{x}\right)$ on $(0,\infty)$. Associated with a couple (X,Y) of positive r.v.'s consider

$$(U,V) := \left(\frac{1}{e^{X+Y}-1},\ \frac{1}{e^X - 1} - \frac{1}{e^{X+Y}-1}\right). \tag{2.1}$$

Then

$$(X,Y) = \left(\log\left(1 + \frac{1}{U+V}\right),\ \log\left(1+\frac{1}{U}\right) - \log\left(1+\frac{1}{U+V}\right)\right). \tag{2.2}$$

Theorem 2.1 *Consider positive random variables X and Y. Define U and V as above. Assume that X and Y have densities p_X and p_Y respectively, that p_X and p_Y are positive and that $\log p_X$ and $\log p_Y$ are locally integrable.*

If X and Y are independent and if U and V are independent, then

$$p_Y(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(1-e^{-y})^{b-1}e^{-ay}\mathbf{1}_{\{y>0\}} \tag{2.3}$$

$$p_X(x) = \alpha(a+b,c,-a)e^{-(a+b)x}(1-e^{-x})^{-b-1} \exp\left(-c\frac{e^{-x}}{1-e^{-x}}\right)\mathbf{1}_{\{x>0\}}. \tag{2.4}$$

where a, b and c are constants such that $a,b,c > 0$ and $\alpha(a+b,c,-a)$ is the constant defined in Equation (1.1).

The laws of U and V are respectively $K^{(2)}(a,b,c)$ and $\gamma(b,c)$.

3 Proof of Theorem 2.1

Let us first prove the following technical lemma.

Lemma 3.1 *Let $\theta : (0,\infty) \to (0,\infty)$ be a locally integrable function. Consider $\phi = (\phi_1, \phi_2) : (0,\infty)^2 \to (0,\infty)^2$ and $\psi = (\psi_1, \psi_2)$, such that ϕ and ψ are C^1 and*

$$\phi \circ \psi(x,y) = (x,y), \quad x,y > 0 \tag{3.5}$$

and such that, for each fixed $y > 0$, the map $\psi_{1,y} : x \mapsto \psi_1(x,y)$ is C^1 and bijective on $(0,\infty)$. Suppose the existence of a function G such that

$$G(s) - G(t) = \theta\big(\phi_1(s,t)\big) - \theta\big(\phi_2(s,t)\big), \quad s,t > 0. \tag{3.6}$$

Then G and θ are C^1.

Proof: By (3.5), $G\big(\psi_1(x,y)\big) - G\big(\psi_2(x,y)\big) = \theta(x) - \theta(y), \quad x,y > 0$. Consider any $x_0, x_1 > 0$. Since θ is locally integrable, G is locally integrable and we have :

$$\int_{x_0}^{x_1} G\big(\psi_1(x,y)\big)\,dx - \int_{x_0}^{x_1} G\big(\psi_2(x,y)\big)\,dx = \int_{x_0}^{x_1} \theta(x)\,dx - (x_1 - x_0)\theta(y). \tag{3.7}$$

After the change of variable $s = \psi_1(x,y)$ in the first integral and the change of variable $t = \psi_2(x,y)$ in the second integral, (3.7) gives

$$\begin{aligned} &\int_{\psi_1(x_0,y)}^{\psi_1(x_1,y)} G(s)\big(\psi_{1,y}^{-1}\big)'(s)\,ds \quad - \quad \int_{\psi_2(x_0,y)}^{\psi_2(x_1,y)} G(t)\big(\psi_{2,y}^{-1}\big)'(t)\,dt \\ &= \int_{x_0}^{x_1} \theta(x)\,dx - (x_1 - x_0)\theta(y) \quad . \end{aligned} \tag{3.8}$$

Since the left-hand side of (3.8) is continuous in y (because ψ is C^1), the function θ is continuous. Using (3.6) we deduce from the continuity of θ that G is continuous (because ϕ is continuous). Since G is continuous, the left-hand side of (3.8) is a C^1 function in y. Consequently, θ is C^1. We deduce, again from (3.6), that G is C^1. □

Suppose that X and Y are independent and fulfill the assumptions of Theorem 2.1, and that U and V are independent. Let us prove that X, Y, U and V have the prescribed densities.

X and Y being independent, as U and V, we have

$$p_U(u)p_V(v) = p_X(g(u+v))\,p_Y(g(u) - g(u+v))\,g'(u+v)g'(u). \tag{3.9}$$

Using $(X,Y) = (g(U+V), g(U) - g(U+V))$ and reasoning as for (3.9) we get:

$$p_X(z)p_Y(w) = p_U(f(z+w))\,p_V(f(z) - f(z+w))\,f'(z+w)f'(z), \quad z,w > 0. \tag{3.10}$$

Recall that, according to (3.9), we also have

$$p_U(z)p_V(w) = p_X(g(z+w))\,p_Y(g(z) - g(z+w))\,g'(z+w)g'(z), \quad z,w > 0. \tag{3.11}$$

Lemma 3.2 *Consider* $h := -\frac{p_X}{p_Y f'}$, $k := -\frac{p_U}{p_V g'}$, $H(r) := \log\Big(h\Big(\log\big(1+\frac{2}{r}\big)\Big)\Big)$, $r>0$, $F := \log k$, $\alpha := \log p_V$, $\beta := \log p_Y$.

Then the following functional equations hold :

$$H(s) - H(t) = \alpha\left(\frac{s(s+2)}{2(s+t+2)}\right) - \alpha\left(\frac{t(t+2)}{2(s+t+2)}\right), \quad s,\, t>0. \tag{3.12}$$

$$F(u) - F(v) = \beta\big(g(u) - g(u+v)\big) - \beta\big(g(v) - g(u+v)\big), \quad u,\, v>0. \tag{3.13}$$

Proof of Lemma 3.2. From (3.10) we have

$$p_X(w)p_Y(z) = p_U\big(f(z+w)\big)\, p_V\big(f(w) - f(z+w)\big)\, f'(z+w)f'(w), \quad z,w>0. \tag{3.14}$$

Dividing the left-hand side of (3.10) by the left-hand side of (3.14) and taking the logarithm, we have

$$\log h(z) - \log h(w) = \log\Big(p_V\big(f(z) - f(z+w)\big)\Big) - \log\Big(p_V\big(f(w) - f(z+w)\big)\Big). \tag{3.15}$$

Set $z = \log\big(1+\frac{2}{s}\big)$, $w = \log\big(1+\frac{2}{t}\big)$. (Note that $z,w>0 \Leftrightarrow s,t>0$). Then the left-hand side of (3.15) is $H(s)-H(t)$. We have successively

$$f(z) = \frac{1}{e^z-1} = \frac{s}{2}, \quad f(w) = \frac{t}{2}, \quad f(z+w) = \frac{st}{2(s+t+2)}.$$

Therefore, $f(z) - f(z+w) = \frac{s(s+2)}{2(s+t+2)}$, $\quad f(w) - f(z+w) = \frac{t(t+2)}{2(s+t+2)}$. Thus, the right-hand side of (3.15) is $\alpha\left(\frac{s(s+2)}{2(s+t+2)}\right) - \alpha\left(\frac{t(t+2)}{2(s+t+2)}\right)$ and we get (3.12). To get (3.13) we write, using (3.11),

$$p_U(w)p_V(z) = p_X\big(g(z+w)\big)\, p_Y\big(g(w) - g(z+w)\big)\, g'(z+w)g'(w), \quad z,w>0, \tag{3.16}$$

we divide the two sides of (3.11) by those of (3.16) and take the logarithm. □

We will need the following lemma, which proof comes easily from the definitions.

Lemma 3.3 *If* $\log(p_X)$ *and* $\log(p_Y)$ *are locally integrable over* $]0,\infty[$, *then* $\log(p_U)$, $\log(p_V)$, H *and* F *are locally integrable.*

Solution of the first functional equation (3.12)

By Lemma 3.1 applied to

$$\phi(s,t) = \left(\frac{s(s+2)}{2(s+t+2)},\, \frac{t(t+2)}{2(s+t+2)}\right), \;\; \psi(x,y) = (2x{-}1{+}\sqrt{4xy+1},\, 2y{-}1{+}\sqrt{4xy+1}),$$

the functions H and α are C^1. Let us differentiate (3.12) with respect to s. With the notation $\hat{s} := \frac{s(s+2)}{2(s+t+2)}$, $\hat{t} := \frac{t(t+2)}{2(s+t+2)}$ we get

$$H'(s) = \alpha'(\hat{s}) \left[\frac{s+1}{s+t+2} - \frac{s(s+2)}{2(s+t+2)^2} \right] + \alpha'(\hat{t}) \frac{t(t+2)}{2(s+t+2)^2}. \tag{3.17}$$

a) Let $t \to 0$ in (3.17). We have $\hat{s} \to s/2$ and $\hat{t} \to 0$ as $t \to 0$. Writing

$$\alpha'(\hat{t}) \frac{t(t+2)}{2(s+t+2)^2} = \hat{t}\alpha'(\hat{t}) \frac{1}{s+t+2}$$

we deduce from (3.17) that $M := \lim_{u \to 0} u\alpha'(u)$ exists and that

$$H'(s) = \alpha'(\frac{s}{2}) \left[\frac{s+1}{s+2} - \frac{s}{2(s+2)} \right] + \frac{M}{s+2},$$

i.e.

$$H'(s) = \frac{1}{2}\alpha'\big(\frac{s}{2}\big) + \frac{M}{s+2}. \tag{3.18}$$

b) Let $t \to \infty$ in (3.17). We write

$$\alpha'(\hat{s}) \frac{s+1}{s+t+2} = \big(\hat{s}\alpha'(\hat{s})\big) \frac{2(s+1)}{s(s+2)},$$

$$\alpha'(\hat{s}) \frac{s(s+2)}{2(s+t+2)^2} = \big(\hat{s}\alpha'(\hat{s})\big) \frac{1}{s+t+2}.$$

Let $t \to \infty$ in (3.17). Since $\lim_{t \to \infty} \hat{s} = 0$, we can use the definition of M to obtain the existence of $2N := \lim_{u \to \infty} \alpha'(u)$ and the fact that $H'(s) = \frac{2M(s+1)}{s(s+2)} + N$ which gives

$$H'(s) = N + \frac{M}{s} + \frac{M}{s+2}, \quad s > 0. \tag{3.19}$$

By (3.18) we get $\alpha'(s) = 2N + \frac{M}{s}$. This implies the existence of a constant L such that

$$\alpha(s) = 2Ns + M \log s + L, \quad s > 0. \tag{3.20}$$

Since $\alpha = \log p_V$, we get that V follow the gamma law with density function

$$p_V(s) = e^{2Ns} s^M e^L, \quad s > 0. \tag{3.21}$$

Solution of the second functional equation (3.13)

By Lemma 3.1 with $\phi(u, v) = (g(u) - g(u+v),\, g(v) - g(u+v))$, the local integrability of $\beta = \log(p_Y)$ implies that the functions F and β are C^1. Let us differentiate (3.13) with respect to u :

$$F'(u) = (g'(u) - g'(u+v))\beta'(g(u) - g(u+v)) + g'(u+v)\beta'\big(g(v) - g(u+v)\big), \quad u,\, v > 0. \tag{3.22}$$

a) Let $v \to \infty$ in (3.22). Then calculations show that $A := \lim_{x\to 0} x\beta'(x)$ exists and

$$F'(u) = g'(u)\beta'(g(u)) - \frac{A}{u}, \quad u > 0. \tag{3.23}$$

b) Let $v \to 0$ in (3.22). Then we see that $B := \lim_{x\to\infty} \beta'(x)$ exists and

$$F'(u) = -\frac{1+2u}{u(1+u)}A + Bg'(u), \quad u > 0. \tag{3.24}$$

c) Computing β. From (3.23) and (3.24) we get $g'(u)\beta'(g(u)) = -\frac{A}{1+u} + Bg'(u)$, i.e.

$$\beta'(g(u)) = -\frac{A}{(1+u)g'(u)} + B, \quad u > 0. \tag{3.25}$$

By integration, this gives $\beta(v) = A\log(1-e^{-v}) + Bv + C$, where C is a constant. Since $\beta = \log p_Y$ we get

$$p_Y(v) = e^C(1-e^{-v})^A e^{Bv}, \quad v > 0 \tag{3.26}$$

which is the desired result for the density of Y (we have $B < 0$ and $A > 0$).

Let us compute the density of U. By (3.24) we have $F'(u) = -(A+B)\frac{1}{u} + (B-A)\frac{1}{u+1}$. As a consequence, $F(u) = -(A+B)\log u + (B-A)\log(u+1) + J$, where J is a constant. Thus, $k(u) = e^{F(u)} = u^{-A-B}(u+1)^{B-A}e^J$. It follows from the definition of k and from (3.21) that, for all $u > 0$,

$$p_U(u) = -k(u)p_V(u)g'(u) = e^{L+J}u^{-A-B+M-1}(u+1)^{B-A-1}e^{2Nu} \quad u > 0, \tag{3.27}$$

and this proves that U follows a Kummer distribution.

Let us compute the density of X. Integrating (3.19) leads to

$$H(s) = Ns + M\log s + M\log(s+2) + I, \quad s > 0, \tag{3.28}$$

where I is a constant. Since $H(s) := \log\Big(h\Big(\log\big(1+\frac{2}{s}\big)\Big)\Big)$ we have

$$h\left(\log\big(1+\frac{2}{s}\big)\right) = e^I e^{Ns}s^M(s+2)^M, \quad s > 0. \tag{3.29}$$

Setting $x = \log(1+\frac{2}{s})$ we have $s = \frac{2}{e^x-1}$ and

$$h(x) = e^I \exp\left(\frac{2N}{e^x-1}\right)\left(\frac{2}{e^x-1}\right)^M\left(\frac{2}{e^x-1}+2\right)^M, \quad x > 0. \tag{3.30}$$

As a result, by the definition of h and (3.26) we obtain

$$p_X(x) = e^{C+I}2^{2M}\exp\left(\frac{2N}{e^x-1}\right)e^{(B-M-1)x}(1-e^{-x})^{A-2M-2}, \quad x > 0, \tag{3.31}$$

which his the expected result. Plugging the expressions of the four densities (3.21), (3.26), (3.27) and (3.31) in (3.10) we obtain $A = M$. This ends the proof of the theorem, with $A = b-1$, $B = -a$, $M = b-1$. □

References

[1] Koudou, E. and Vallois, P. (2010). Independence properties of the Matsumoto-Yor type. To appear in *Bernoulli*.

[2] Letac, G. and Wesolowski, J. (2000). An independence property for the product of GIG and gamma laws. *Ann. Prob.* **28**, 1371-1383.

[3] Matsumoto, H. and Yor, M. (2001). An analogue of Pitman's $2M - X$ theorem for exponential Wiener functional, Part II: the role of the generalized inverse Gaussian laws. *Nagoya Math. J.* **162**, 65-86.

[4] Ng, K. W. and Kotz, S. (1995). Kummer-Gamma and Kummer-Beta univariate and multivariate distributions. *Research report*, Department of Statistics, The University of Hong Kong, Hong Kong.

[5] Nagar and Gupta (2002). Matrix-variate Kummer-beta distributions. *J. Austral. Math. Soc.* **73**, 11-25.

[6] Wesolowski, J. (2002). On a functional equation related to the Matsumoto-Yor property. *Aequationes Math.* **63**, 245-250.

DISTRIBUTION OF THE DETERMINANT OF A SAMPLE CORRELATION MATRIX IN and APPLICATIONS

Thu PHAM-GIA,
Département de statistique, Université de Moncton, Moncton, NB Canada E1A 3E9

Mots-Clés : Statistique mathématique, Fiabilité, Matrice de corrélation, Meijer
Mathematical statistics, Reliability, Correlation matrix, Meijer.

Summary: The correlation matrix in multivariate statistical analysis is often used in applications. However, its exact distribution is only known when the population correlation matrix is the Identity, except for the bivariate case where several results are available. Also, the distribution of its determinant is unknown. We study here the latter distribution, using Meijer functions G, which are computable, and apply it to the concept of dependence for a system. For the general case where the population correlation matrix is not Identity a simulation study provides some results.

Résumé : La matrice de corrélation en statistique multidimensionnelle est fréquemment utilisée dans les applications. Cependant sa distribution exacte est seulement connue quand la matrice de corrélation de la population est l'Identité, sauf dans le cas d'une distribution bivariée. Aussi, la distribution de son déterminant est inconnue. Dans cette communication nous étudions cette dernière distribution avec les fonctions spéciales de Meijer G, qui sont à présent calculables. Nous l'appliquons ensuite au concept de dépendance d'un système. Pour le cas général quand la matrice de corrélation de la population n'est pas l'Identité une étude par simulation procure quelques résultats.

1. INTRODUCTION

There are several uses of the determinant $|\mathbf{R}|$ of the sample correlation matrix $\mathbf{R}$ in multivariate statistical analysis (Reddon et al.(1985)) and in Signal processing (Leshem and Van der Veenl (2001)). We apply it here to Reliability Theory.

2. THE POPULATION AND SAMPLE CORRELATION MATRICES

2.1 Covariance and correlation matrices: Let us consider a random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, of the form of a *(p x p)* symmetric positive definite random matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12}... & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \\ \sigma_{p1} & & \sigma_{pp} \end{bmatrix}$ of pairwise covariances between components in the matrix.

We obtain the population correlation matrix $\boldsymbol{\Lambda}$ by dividing each σ_{ij} by $\sqrt{\sigma_{ii}\sigma_{jj}}$. Then $\boldsymbol{\Lambda} = \left[\rho_{ij}\right] = \mathbf{D}_{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{D}_{\Sigma}^{-\frac{1}{2}}$, where $\mathbf{D}_{\Sigma} = Diag(\sigma_{11},...,\sigma_{pp})$, is symmetrical, with diagonals $\rho_{ii} = 1$, $i = 1,..., p$, i.e $\boldsymbol{\Lambda} = \begin{bmatrix} 1 & \rho_{12}... & \rho_{1p} \\ \rho_{21} & 1 & \\ \rho_{p1} & & 1 \end{bmatrix}$. For a sample of size *n* of observations from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the sample mean $\overline{\mathbf{X}} = \sum_{\alpha=1}^{n} \mathbf{X}_{\alpha} / n$ and the "adjusted sample covariance" matrix

$\mathbf{S} = \sum_{\alpha=1}^{n} \left(\mathbf{x}_{\alpha} - \overline{\mathbf{x}}\right)\left(\mathbf{x}_{\alpha} - \overline{\mathbf{x}}\right)^{T}$ (or matrix of sums of squares and products, $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12}... & s_{1p} \\ s_{21} & s_{22} & \\ s_{p1} & & s_{pp} \end{bmatrix}$)

are independent, with the latter having a Wishart distribution $\mathbf{W}_p(n-1,\boldsymbol{\Sigma})$. The sample correlation matrix $\mathbf{R}=\begin{bmatrix} 1 & r_{12}... & r_{1p} \\ r_{21} & 1 & \\ r_{p1} & & 1 \end{bmatrix}$ is obtained from the $\mathbf{S}$ by using: $r_{ij}=s_{ij}/\sqrt{s_{ii}s_{jj}}$, or $\mathbf{R}=\mathbf{D}_s^{-\frac{1}{2}}\mathbf{S}\mathbf{D}_s^{-\frac{1}{2}}$, where $\mathbf{D}_S=Diag(s_{11},s_{22},...,s_{pp})$. We also have : $|\boldsymbol{\Sigma}|=|\boldsymbol{\Lambda}|\prod_{i=1}^{p}\sigma_{ii}$, and similarly, $|\mathbf{S}|=|\mathbf{R}|\prod_{i=1}^{p}s_{ii}$. $\boldsymbol{\Sigma}=diag(\sigma_{11},...\sigma_{pp})\Leftrightarrow(\boldsymbol{\Lambda}=\mathbf{I})$. When considering $S^*=\mathbf{S}/(n-1)\sim W_p((n-1),\boldsymbol{\Sigma}/(n-1))$, we have $\left\{(s_i^*)^2=s_{ii}/(n-1)\right\}, i=1,...,p$, between the two diagonal elements, but exactly the same sample correlation matrix.

R can always be defined from **S**, the reverse is not true. In the bivariate case, Hotelling's expression:

$$f(r)=\frac{n-2}{\sqrt{2\pi}}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}(1-\rho^2)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}(1-\rho r)^{-(n+\frac{1}{2})}{}_2F_1(1/2,1/2;n+1/2;\frac{1+\rho r}{2}),-1\le r\le 1 \quad (1)$$

where ${}_2F_1(\alpha,\beta;\lambda;x)$ is Gauss hypergeometric function, clearly shows f(*r*) depends on ρ and n only.

For the bivariate case, we have the expression of the null distribution of r when $\rho\neq 0$, and hence inference procedures for ρ, the same is not true for |R| when $\boldsymbol{\Lambda}\neq\mathbf{I}$ and $|\boldsymbol{\Lambda}|$ can only be estimated via simulation. Since r_{ij} are biased estimators of σ_{ij}, and Olkin and Pratt (1958) have suggested using the modified estimator $r_{ij}\left\{1+\frac{1-r_{ij}^{\ 2}}{2(n-4)}\right\}$. Pham-Gia and Turkkan (2010) give a closed formed expression for the density of R in this case, conditional to the density of the variances, which reduces to (2) when $\boldsymbol{\Lambda}=\mathbf{I}$.

3. CASE OF THE POPULATION CORRELATION MATRIX BEING IDENTITY

3.1 Density of R: We have:

Hence, $f(\mathbf{R})=\frac{\left[\boldsymbol{\Gamma}(\frac{n-1}{2})\right]^p|\mathbf{R}|^{\frac{n-p-2}{2}}}{\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\boldsymbol{\Gamma}(\frac{n-i}{2})}=\frac{\left[\boldsymbol{\Gamma}(\frac{n-1}{2})\right]^p|\mathbf{R}|^{\frac{n-p-2}{2}}}{\boldsymbol{\Gamma}_p(\frac{n-1}{2})}$ (2)

3.2 **Density of $|\mathbf{R}|$**: Meijer **G**-functions is used here. Let us recall that (Pham-Gia (2008) that $\mathbf{G}(x)=\mathbf{G}^{m\,r}_{p\,q}\left[x\middle|\begin{matrix}a_1,...,a_p\\ b_1,...,b_q\end{matrix}\right]$ is the integral along the complex contour L of a rational expression of Gamma functions, i.e.

$$\mathbf{G}^{\,m\,r}_{p\,q}\left[x\left|\begin{matrix}a_1,...,a_p\\ b_1,...,b_p\end{matrix}\right.\right]=\frac{1}{2\pi i}\int_L \frac{\prod_{j=1}^{m}\Gamma\left(b_j-s\right)\prod_{j=1}^{r}\Gamma\left(1-a_j+s\right)}{\prod_{j=m+1}^{q}\Gamma\left(1-b_j+s\right)\prod_{j=r+1}^{p}\Gamma\left(a_j-s\right)}x^s ds\,.$$

From (2) the moments of order t of $|\mathbf{R}|$ can be seen as a product of moments of order t of independent beta variables. Upon identification with these products, we can see that $|\mathbf{R}| \sim X_1...X_{p-1}$, with $X_i \sim \beta(\frac{n-j}{2}, \frac{j-1}{2})$, with $j = 2,3.,.., p$.

Since , using Pham-Gia (2008), the product of k independent betas, $\beta(\alpha^{(i)}, \beta^{(i)})$, has as density

$$\prod_{i=1}^{k}\frac{\Gamma(\alpha^{(i)}+\beta^{(i)})}{\Gamma(\alpha^{(i)})}\mathbf{G}^{\,k\quad 0}_{k\quad k}\left[y\;\middle|\;\begin{matrix}\alpha^{(1)}+\beta^{(1)}-1,.....,\alpha^{(k)}+\beta^{(k)}-1\\ \alpha^{(1)}-1,.....,\alpha^{(k)}-1\end{matrix}\right],0\le y\le 1, \qquad (3)$$

we have here, the density of $|\mathbf{R}|$ as:

$$g(r;n,p)=\prod_{i=1}^{p-1}\frac{\Gamma(\alpha^{(i)}+\beta^{(i)})}{\Gamma(\alpha^{(i)})}\mathbf{G}^{\,p-1\quad 0}_{p-1\quad p-1}\left[\begin{matrix}\alpha^{(1)}+\beta^{(1)}-1,.....,\alpha^{(p-1)}+\beta^{(p-1)}-1\\ \alpha^{(1)}-1,.....,\alpha^{(p-1)}-1\end{matrix}\right],0\le r\le 1,$$

where $\alpha^{(1)}=\frac{n-2}{2},\beta^{(1)}=\frac{1}{2},\ \alpha^{(2)}=\frac{n-3}{2},\beta^{(1)}=\frac{2}{2}$ and $\alpha^{(p-1)}=\frac{n-p}{2},\beta^{(p-1)}=\frac{p-1}{2}$. Hence,

$$g(u\;;n,p)=\frac{\left[\Gamma(\frac{n-1}{2})\right]^{p-1}}{\Gamma(\frac{n-2}{2})...\Gamma(\frac{n-p}{2})}\mathbf{G}^{\,p-1\quad 0}_{p-1\quad p-1}\left[u\;\middle|\;\begin{matrix}\frac{n-3}{2},.....,\frac{n-3}{2}\\ \frac{n-4}{2},.....,\frac{n-(p+2)}{2}\end{matrix}\right],0\le u\le 1 \qquad (4)$$

Fig.1 ($p=4, n=8$), gives percentiles of $|\mathbf{R}|$, determined numerically, with the 2.5th and 97.5th percentiles being 0.04697 and 0.7719 respectively.

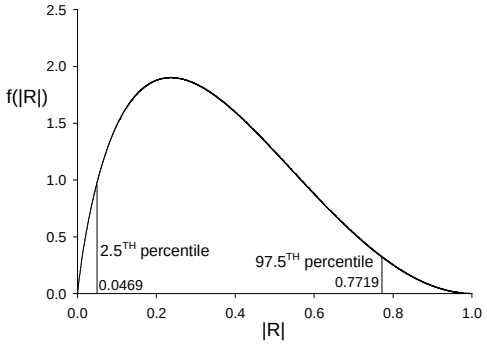


Fig. 1
Density of $|\mathbf{R}|$ for $n = 8$, $p = 4$

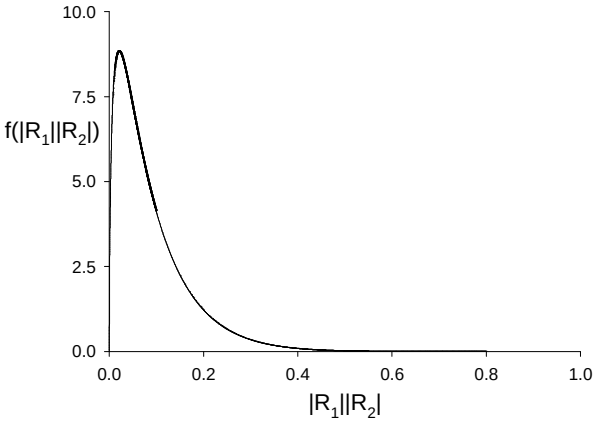


Fig. 2
Density of product of 2 independent correlation determinants, $n_1 = 8$, $p_1 = 4$ $n_2 = 10$, $p_2 = 5$

REMARK : Gupta and Rathie (1983)) gave an equivalent expression. From (3), asymptotically, we have $-n\log_e|\mathbf{R}|\sim\chi^2_{p(p-1)/2}$ (Muirhead (1982, p. 151)) and approximate inferences on ρ can be made.

3.3 Density of the product $|\mathbf{R}_1|.|\mathbf{R}_2|$:

THEOREM 1: Let $\left\{\mathbf{X}_1^{(1)},\mathbf{X}_2^{(1)},...,\mathbf{X}_{n_1}^{(1)}\right\}$ and $\left\{\mathbf{X}_1^{(2)},\mathbf{X}_2^{(2)},...,\mathbf{X}_{n_2}^{(2)}\right\}$ be two independent random samples from $N_{p_1}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $N_{p_2}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ respectively, both with zero population correlation coefficients. Then the determinant $|\mathbf{R}|$ of the product $\mathbf{R}=\mathbf{R}_1\mathbf{R}_2$ of the two correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ has density:

$$g(v;n_1,n_2,p)=A\mathbf{G}^{p-2\quad 0}_{p-2\quad p-2}\left[v\,\middle|\,\begin{matrix}\frac{n_1-3}{2},.....,\frac{n_1-3}{2},\frac{n_2-3}{2},.....,\frac{n_2-3}{2}\\ \frac{n_1-4}{2},.....,\frac{n_1-(p_1+2)}{2},\frac{n_2-4}{2},.....,\frac{n_2-(p_2+2)}{2}\end{matrix}\right],0\le v\le 1, \quad (5)$$

where $A=\prod_{i=1}^{2}\left\{\left[\boldsymbol{\Gamma}(\frac{n_i-1}{2})\right]^{p_i-1}/\left(\boldsymbol{\Gamma}(\frac{n_i-2}{2})...\boldsymbol{\Gamma}(\frac{n_i-p_i}{2})\right)\right\}$ and $p=p_1+p_2$.

PROOF: Applying Pham-Gia's (2008) approach we obtain (5) from (4). **QED**

We can similarly derive the density of the ratio $|\mathbf{R}_1|/|\mathbf{R}_2|$ in terms of G-functions**.**

.

4. APPLICATION IN RELIABILITY: DEPENDENCE WITHIN A SYSTEM

4.1 We approach this dependence concept here, by way of the correlation matrix, that would reflect the overall degree of dependence. It is defined as $\delta(\vartheta)=(1-|\boldsymbol{\Lambda}|)^{1/p}$, (6) with $0\le\delta(\vartheta)\le 1$. The measure of independence within the system is then $1-(1-|\boldsymbol{\Lambda}|)^{1/p}$, estimated by $1-(1-|\hat{\boldsymbol{\Lambda}}|)^{1/p}$, where $\hat{\boldsymbol{\Lambda}}$ is a point estimation of $\boldsymbol{\Lambda}$ based on **R,** the sample correlation matrix . In the case $\boldsymbol{\Lambda}=\mathbf{I}$, as seen above, we can take $\hat{\boldsymbol{\Lambda}}$ = E(|**R**|) , which can be computed.

THEOREM 2: Let the fully statistically independent system ϑ have p components with a joint normal distribution with $\boldsymbol{\Lambda}=\mathbf{I}$, with $p\ge 2$. Then the distribution of the sample coefficient of inner dependence $d(\vartheta)$ is :

$$f(d;n,p)=\frac{\left[\boldsymbol{\Gamma}(\frac{n-1}{2})\right]^{p-1}}{\boldsymbol{\Gamma}(\frac{n-2}{2})...\boldsymbol{\Gamma}(\frac{n-p}{2})}\,p(1-d)^{p-1}\mathbf{G}^{p-1\quad 0}_{p-1\quad p-1}\left[1-(1-d)^p\,\middle|\,\begin{matrix}\frac{n-3}{2},.....,\frac{n-3}{2}\\ \frac{n-4}{2},.....,\frac{n-(p+2)}{2}\end{matrix}\right],0\le d\le 1 \quad (7)$$

The case $\boldsymbol{\Lambda}\ne\mathbf{I}$ will be treated by simulation in section 5. We define a *p*-component normal system to be fully statistically independent when the *p(p-1)/2* correlation coefficients ρ_{ij} of its components are all zero.

4.2 THE BIVARIATE NORMAL: Here we have a complete solution for all cases. We have $|\mathbf{R}|=1-r^2$, and when $\rho=0$, we have , the density of $|\mathbf{R}|$ as

$\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})}\mathbf{G}^{1\ 0}_{1\ 1}\left[u\,|\,\begin{matrix}\frac{n-3}{2}\\ \frac{n-4}{2}\end{matrix}\right]$, $0\le u\le 1$, which is the **G**-function of $\beta(\frac{n-2}{2},\frac{1}{2})$, Hence, the distribution of r^2 is: $\beta(1/2,(n-2)/2)$, $0\le r^2\le 1$, and the density of r is

$f(r)=(1-r^2)^{(n-4)/2}/B(1/2,(n-2)/2)$, for $-1\le r\le 1$. (8).

We then have $|\mathbf{\Lambda}|$=1- ρ^2, and $\delta(\vartheta)=|\rho|$. The associated sample measure being $d(\vartheta)=|r|$.

THEOREM 3: For the bivariate (two-component) case, we have:

1) For a fully independent two-component binormal ($\rho=0$), for $0\le d\le 1$:

$f(n,d)=2(1-d^2)^{(n-4)/2}/B(1/2,(n-2)/2)$, $0\le d\le 1$, (9)

2) For a non fully independent two-component binormal system($\rho\neq 0$), for

$$0\le d\le 1:\ f(n,d)=A.(1-d^2)^{(n-4)/2}\left\{\begin{matrix}(1-\rho d)^{-(n-\frac{3}{2})}{}_2F_1(1/2,1/2;(2n-1)/2,(1+\rho d)/2)+\\ (1+\rho d)^{-(n-\frac{3}{2})}{}_2F_1(1/2,1/2;(2n-1)/2,(1-\rho d)/2)\end{matrix}\right\} \quad (10)$$

with $A=\frac{(n-2)\Gamma(n-1)(1-\rho^2)^{(n-1/)2}}{\Gamma(n-(1/2))(2\pi)^{1/2}}$

PROOF: 1) For $\rho_{i,j}=0$, the density of $d(\vartheta)$, as given by (9), is obtained from (8) by a change of variable. Expression (10) is obtained from (1) by the change of variable, $d=|r|$.

QED.

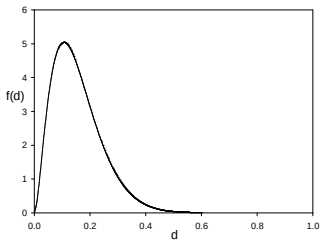


Fig. 3
Density of sample measure of dependence binormal distribution($\rho=0$, n =10, p = 4)

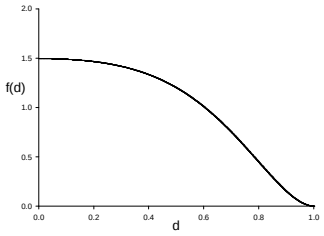


Fig. 4
Density of $d(\vartheta)$ for a system with system dependence, $d(\vartheta)$ (n=8, $\rho=0.25$)

Numerical computations give $E(d)=0.17665$ for the first case. Estimation of $\delta(\vartheta)$ from $d(\vartheta)$ now follows the same principles as ρ from r.

5. SOME RESULTS FOR THE MULTINORMAL CASE (**p** ≥ 3 and $\mathbf{\Lambda}\neq\mathbf{I}$).

5.1 Simulations related to R : Starting from a normal model $\mathbf{X}\sim N_p(\boldsymbol{\mu},\mathbf{\Lambda})$, $\mathbf{\Lambda}\neq\mathbf{I}$, simulation can be used. We start from the *(4 ×4)* population covariance matrix taken from Pham-Gia, Turkkan & Vovan(2008).

$\mathbf{\Omega} = \begin{bmatrix} 0.122 & 0.097 & 0.016 & 0.010 \\ & 0.140 & 0.011 & 0.009 \\ & & 0.030 & 0.006 \\ & & & 0.011 \end{bmatrix}$, $\mathbf{\Lambda} = \begin{bmatrix} 1 & 0.7425 & 0.2672 & 0.2781 \\ & 1 & 0.1777 & 0.2328 \\ & & 1 & 0.3316 \\ & & & 1 \end{bmatrix}$, where all $\rho_{ij} \neq 0$. We generate 10000 samples of 100 observations each from $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which give 10000 values of the covariance matrix S, which, in turn, give matrix values for **R**, scalar values for $|\mathbf{R}|$.

5.2 The simulated distribution of $|\mathbf{R}|$ is given by Fig 6. Replacing the values r_{ij} by the corrected value $r_{ij}\left\{1+\frac{1-r_{ij}^{\,2}}{2(n-4)}\right\}$, the unbiased estimator of ρ_{ij}, we obtain $|\hat{\mathbf{R}}|$. also given in Fig. 6. The approximate density of $|\hat{\mathbf{R}}|$ has higher mean and median. The two variation intervals are very small. Taking the mean of the distribution of $|\hat{\mathbf{R}}|$, we have: E($|\hat{\mathbf{R}}|$= 0.84 x 10^{-5} from which an estimate of $|\mathbf{\Lambda}|$ can be made.

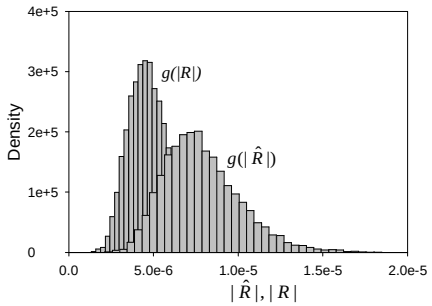


Fig. 6 Simulated densities of $|\mathbf{R}|$ and $|\hat{\mathbf{R}}|$

BIBLIOGRAPHY

[1]Gupta, A.K. and Rathie, P.N., On the Distribution of the Determinant of Sample Correlation Matrix from Multivariate Gaussian population, *Metrika*, 41, 1983,43-56.

[2]Hotelling, H., New Light on the Correlation Coefficient and its Transform, *J. R. Statist. Soc.*, B, 15, 1953, 193.

[3]Leshem, A. and Van der Veen, A-J., Multichannel Detection of Gaussian Signals with Uncalibrated Receivers, *IEEE Signal Processing Letters*, 8. 4. 2001, 120-122.

[4]Muirhead, R. J., *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[5]Pham-Gia, T., Exact Distribution of the Generalized Wilks's Statistic and Applications, *Journ. of Multivariate Analysis*, 2008, 1698-1716.

[6]Pham-Gia, T. and Turkkan, N., Distribution of the Sample Correlation Matrix, *Statistics*, submitted , 2010.

[7]Reddon, John R., Jackson, D.N. and Schopflocher, D., Distribution of the Determinant of the Sample Correlation Matrix: Monte Carlo Type One Error Rates, *J. of Ed. Stat.*, 10, 4, 1985, 384-88.

Echantillonnage conditionnel

Broniatowski Michel & Caron Virgile

Laboratoire de Statistique Théorique et Appliquée (LSTA)
UPMC - Paris 6, Tour 15-25, 2ème étage
4 place Jussieu, 75252 Paris cedex 05

Abstract

Nous présentons un schéma d'approximation de la densité d'un vecteur aléatoire $(\mathbf{X}_1, ..., \mathbf{X}_k)$ conditionné à la valeur observée de $\bar{X}_n = \frac{1}{n}(\mathbf{X}_1 + ... + \mathbf{X}_n)$ lorsque $k/n \to 1$; on traite également le cas où le conditionnement est de la forme $\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i)$. Les conditionnements condidérés sont supposés "normaux", c.a.d. dans la gamme des valeurs impliquées par la loi du log-itéré. La qualité de l'approximation est étudiée et un algorithme explicite fournit la valeur maximale de k, en fonction de n, pour laquelle une précision relative fixée est atteinte. L'approximation proposée permet la simulation d'échantillons $(Y_1, ..., Y_k)$ sous le conditionnement induit par l'observation $\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i)$. Nous utilisons cet échantillonnage conditionnel pour la tabulation de tests relatifs à un paramètre d'interêt sous un paramètre de nuisance, lorsque le conditionnement est définie par la valeur prise par une statistique exhaustive pour ce paramètre de nuisance. Ces méthodes sont aussi applicables pour l'amélioration d'estimateurs dans le cadre du théorème de Rao-Blackwell.

We study the density of a random vector $(\mathbf{X}_1, ..., \mathbf{X}_k)$ conditionned upon the observed value of $\bar{X}_n = \frac{1}{n}(\mathbf{X}_1 + ... + \mathbf{X}_n)$. The case where the observed value is $\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i)$ is also considered. A sharp approximation scheme is provided in the case when $k/n \to 1$, namely on long runs. The value of observed conditionning statistics is assumed to be in the range provided by the law of the iterated logarithm. An algorithm is provided in order to obtain the maximal value of k, depending on n, which ensures a prescribed relative accuracy for the approximation. The approximate density leads to a simulation scheme for sampling $(Y_1, ..., Y_k)$ under the conditioning $\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i) = s$. This conditional sampling scheme is used in order to state critical point event values for tests in models with nuisance parameters, conditioning on the sufficient statistics for the nuisance. Also the conditional sampling scheme can be used for Rao-Blackwellization of estimators.

Keywords: Statistique mathématique, Schéma d'approximation, Paramètre de nuisance, Simulation, Rao-blackwellisation.

1 Introduction et notations.

Cet exposé traite du comportement asymptotique de la distribution d'une marche aléatoire conditionnée par la valeur finale de sa somme quand le nombre de variables augmente.

Notons $\mathbf{X}_1^n := (\mathbf{X}_1, ..., \mathbf{X}_n)$ n copies indépendantes d'une variable aléatoire réelle $\mathbf{X}$ de densité p sur $\mathbb{R}$ et $\mathbf{U}_1^n := f(\mathbf{X}_1) + ...+f(\mathbf{X}_n)$, où f est une fonction réelle. Nous notons $E[f(\mathbf{X})] = \mu$ et $Var(f(\mathbf{X})) = \sigma^2$. La fonction caractéristique de la variable aléatoire $f(\mathbf{X})$ est supposée appartenir à L^r pour $r \geq 1$. Nous considérons l'approximation en densité du vecteur $\mathbf{X}_1^k = (\mathbf{X}_1, ..., \mathbf{X}_k)$ sur $\mathbb{R}^k$ quand $\mathbf{U}_1^n = n(a_n\sigma + \mu)$ avec $a = a_n$ satisfaisant

$$\limsup_{n\to\infty} \frac{\sqrt{n}a_n}{\sqrt{2\log\log n}} = 1 \tag{1}$$

et $k := k_n$ est une suite d'entiers telles que

$$0 \leq \limsup_{n\to\infty} k/n \leq 1 \tag{2}$$

avec

$$\lim_{n\to\infty} n - k = \infty. \tag{3}$$

Des résultats équivalents quand les $\mathbf{X}_i$ sont des vecteurs aléatoires de $\mathbb{R}^d$ et quand f est une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R}^s$ seront aussi proposés pour d et s plus grand que 1. Le cas $d = 1$ et (1) remplacé par une condition de type grande ou moyenne déviation est traité dans Broniatowski et Caron (2010).

L'événement conditionnant s'écrit donc:

$$\mathcal{E}_n^f := (\mathbf{U}_1^n = n(a_n\sigma + \mu))$$

On suppose que

$$\phi_f(t) := E\exp tf(\mathbf{X}) < \infty$$

pour t dans un voisinage non vide de 0. On définit les fonctions $m_f(t), s_f^2(t)$ et $\mu_{f,3}(t)$ comme la première, deuxième et troisième dérivée de $\log\phi_f(t)$.

Notons $\mathfrak{P}_n^f$ la distribution de $\mathbf{X}_1^n$ conditionné par $\mathcal{E}_n^f$ et $\mathfrak{p}_n^f$ sa densité sur $\mathbb{R}^k$, $k < n$. On notera

$$\mathfrak{p}_n^f\left(\mathbf{X}_1^k = Y_1^k\right) := p(\mathbf{X}_1^k = Y_1^k \big|\, \mathbf{U}_1^n = n(\sigma a + \mu)). = p_{\mathbf{X}_1^k|\mathbf{U}_1^n = n(\sigma a+\mu)}.\left(Y_1^k\right) \tag{4}$$

2 Approximation de la densité conditionnelle.

Nous introduisons une suite positive ϵ_n qui satisfait

$$\lim_{n\to\infty} \epsilon_n\sqrt{n-k} = \infty \tag{E1}$$

$$\lim_{n\to\infty} \epsilon_n(\log n)^2 = 0. \tag{E2}$$

Nous montrerons que $\epsilon_n(\log n)^2$ est la vitesse de convergence de l'approximation.

Notons

$$\pi_f^{\alpha}(x) := \frac{\exp tf(x)}{\phi_f(t)} p_{\mathbf{X}}(x)$$

avec $m_f(t) = \alpha$ et α appartient au support de P_f, la distribution de $f(\mathbf{X})$ supposée absolument continue.

Nous définissons dans la suite une densité $h_{\sigma a+\mu}(y_1^k)$ avec f.d.r. $H_{\sigma a+\mu}$ sur $\mathbb{R}^k$. Notons

$$h_0(y_1|\, y_0) := \pi_f^{\sigma a+\mu}(y_1)$$

où y_0 est arbitraire, et pour $1 \leq i \leq k-1$ définissons $h_i(y_{i+1}|\, y_1^i)$ de manière récursive.

Soit t_i l'unique solution de l'équation

$$m_i := m_f(t_i/\sigma) = \frac{n}{n-i}\left(\sigma a + \mu - \frac{u_1^i}{n}\right) \tag{5}$$

où $u_1^i := f(y_1) + ... + f(y_i)$.

Définissons

$$h_i(y_{i+1}|\, y_1^i) = C_i p_{\mathbf{X}}(y_{i+1}) \mathfrak{n}(\alpha\beta, \alpha, (f(y_{i+1}) - \mu)/\sigma) \tag{6}$$

où C_i est une constante de normalisation et $\mathfrak{n}(\mu, \tau, x)$ est la densité normale de moyenne μ et de variance τ en x. Les paramètres ci-dessus sont définies par:

$$\alpha = \sigma^{-2} s_f^2(t_i/\sigma)(n-i-1) \tag{7}$$

$$\beta = t_i + \frac{\sigma^4 \mu_{f,3}(t_i/\sigma)}{2\sigma^3 s_f^4(t_i/\sigma)(n-i-1)}. \tag{8}$$

On défnit alors

$$h_{\sigma a+\mu}\left(y_1^k\right) := \prod_{i=0}^{k-1} h_i(y_{i+1}|\, y_1^i). \tag{9}$$

Théorème 1 *En supposant (1), (E1) et (E2). Alors (i)*

$$\mathfrak{p}_n^f\left(\mathbf{X}_1^k = Y_1^k\right) = h_{\sigma a+\mu}(Y_1^k)(1 + o_{\mathfrak{P}_n^f}(\epsilon_n (\log n)^2))$$

and (ii)

$$\mathfrak{p}_n^f\left(\mathbf{X}_1^k = Y_1^k\right) = h_{\sigma a+\mu}(Y_1^k)(1 + o_{H_{\sigma a+\mu}}(\epsilon_n (\log n)^2)).$$

Preuve. Voir Broniatowski et Caron (2011). ■

Remarque 2 *On remarque que l'approximation est obtenue sur les trajectoires typiques; (i) sous l'échantillonage conditionnel, (ii) sous l'échantillonnage approximant la loi conditionnelle; c'est ce cas qui est utilisé en statistique.*

Remarque 3 *Quand les $\mathbf{X}_i$ sont i.i.d. de loi normal centrée réduite et que f est la fonction définie sur $\mathbb{R}$ par $f(x) = x$, alors le résultat du théorème d'approximation précédent est vrai sans aucun terme d'erreur pour tout k=1,2,...,n-1.*

3 Choix de k.

La précision de l'approximation est determiné par

$$RE(k) := \sqrt{Var_{h_{\sigma a+\mu}} 1_{D_k}\left(Y_1^k\right) \frac{\mathfrak{p}_n^f\left(Y_1^k\right) - h_{\sigma a+\mu}\left(Y_1^k\right)}{\mathfrak{p}_n^f\left(Y_1^k\right)}} \tag{10}$$

avec $D := D_k \times \mathbb{R}^{n-k}$ où

$$D_k := \{x_1^k : |\frac{h_{\sigma a+\mu}(x_1^k)}{p_n^f(x_1^k)} - 1| < c\epsilon_n(\log n)^2\} \tag{11}$$

où c est une constante indépendante de k et n. D_k est donc l'ensemble portant l'essentiel de la masse de $H_{\sigma a+\mu}$ et $\lim H_{\sigma a+\mu}(D_k) = 1$.

Après developpement, la régle proposée est simple à implémenter et dinne de bons résultats. Soit δ un niveau de précision voulu pour $RE(k)$. Le graphe de $RE(k)$ permet d'obtenir la valeur supérieure de k pour laquelle l'erreur relative est inférieure à δ:

$$k_\delta := \sup\left(k : RE(k) \leq \delta\right).$$

Sans entrer dans la forme explicite de l'approximation de $RE(k)$ (voir Broniatowski et Caron (2011)), les graphes suivants montrent la bonne adéquation de l'approximation pour des grandes valeurs de k (cas exponentiel). Sur la courbe en pointillé, on peut lire la valeur approchée de $RE(k)$ tandis que la courbe pleine représente la vraie valeur de $RE(k)$.

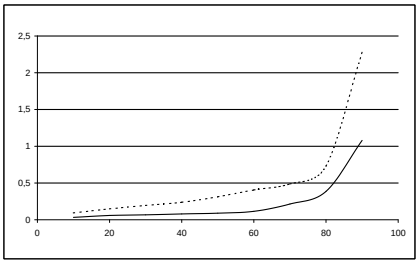


Figure 1: RE(k) en fonction de k pour n=100 et $P(\mathbf{S}_1^n > na) \simeq 10^{-1}$.

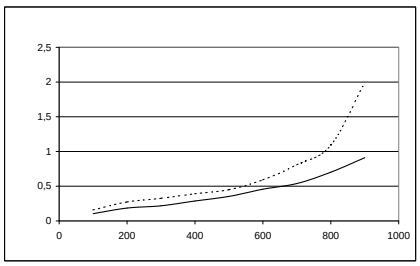


Figure 2: RE(k) en fonction de k pour n=1000 et $P\left(\mathbf{S}_1^n > na\right) \simeq 10^{-1}$.

4 Tabulation d'un test en présence d'un paramètre de nuisance.

On suppose que la distribution de l'échantillon X_1^n est indexée par deux paramètres (θ, η). Le paramètre d'interet est θ et η celui de nuisance. Nous nous interessons à effectuer de l'inférence sur θ conditionnellement à une statistique exhaustive U_1^n pour η. Notre propos s'interesse a la région critique pour une hypothèse simple $\theta = \theta_0$ conditionnée à la valeur observée de $S(X_1^n) = a_n$. Des approches similaires ont été proposés par de nombreux auteurs, c'est aussi le sujet du Chapitre 5 dans Lehman (1986).

L'échantillon observé est supposé réel et i.i.d. d'une famile exponentielle de dimension deux avec paramètres (θ, η):

$$dP_{\theta,\eta}(x) = \exp\left(\theta t(x) + \eta f(x) - K(\theta, \eta)\right) \mathbb{1}_\Omega(x) dx$$

avec Ωest inclus dans $\mathbb{R}$.

A θ_0 fixé, une statistique exhaustive pour le vecteur aléatoire X_1^n est U_1^n pour laquelle pour tout θ

$$dP_\theta(\, x_1^n |\, U_1^n = a_n) = \exp\left(\theta\left(t(x_1) + ... + t(x_n)\right) - K_n(\theta)\right) \prod_{i=1}^{n} \mathbb{1}_\Omega(x_i) dx_1 ... dx_n$$

pour une normalisation $K_n(\theta)$.

On suppose que

$$\lim_{n\to\infty} a_n = \eta_0.$$

Considerons le test de Neyman-Pearson pour $\theta = \theta_0$ versus $\not\theta \neq \theta_0$ conditionnellement à $U_1^n = a_n$ pour la statistique T_1^n avec $T_1^n = \frac{1}{n}\sum_{i=1}^n t(X_i)$, où $t(.)$ telle que $E_{\mathfrak{P}_n}(t(\mathbf{X}))$ est finie. Soit s_α fixé et on définit $0 < \alpha < 1$ le niveau du test telle que

$$P[T_1^n > s_\alpha | S_1^n = n a_n] = \mathfrak{P}_n(T_1^n > s_\alpha) = \alpha. \tag{12}$$

On propose comme estimateur pour α :

$$\hat{\alpha} := \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{(s_\alpha, +\infty)}(T(Y_1^k(l))) \tag{13}$$

où les $Y_i^k(l)$ sont de densité $h_{\sigma a_n + \mu}$ définie dans la section 2 avec $E_{\theta_0}[f(\mathbf{X})] = \mu$ et $Var_{\theta_0}(f(\mathbf{X})) = \sigma^2$.

On suppose, de plus, qu'il existe une fonction $C(x)$ définie sur $\mathbb{R}^+$ qui tends vers 0 à l'infini telle que

$$\liminf_{x\to\infty} \sup_{\lambda \in V(0)} \frac{\exp\{\lambda f(x)\} P[X > x]}{C(x)} > 0 \tag{14}$$

où $V(0)$ est un voisinage fixé de l'origine sur lequel $E[e^{\lambda f(x)}]$ est supposée finie.

Le théorème principal de cette section est le suivant:

Théorème 4 *Soient* $0 < \gamma < 1$ *et* δ_n^L *telle que*

$$\lim_{L\to\infty} \delta_n^L = 0 \tag{15}$$

$$\lim_{L\to\infty} L\delta_n^L = \infty \tag{16}$$

Alors

$$\hat{\alpha} = \alpha + O\left(\left(\frac{n-k}{k}\right)^{\gamma}\right) + O(\epsilon_n(\log n)^2) + o_{H_{\sigma a+\mu}}(\delta_n^L). \tag{17}$$

La figure 3 présente une tabulation de la f.r.d. complémentaire de $\hat{\theta}$ (courbe rouge) pour la loi $\Gamma(\theta, \eta)$ où la statistique exhauxtive pour η est $\frac{1}{n}\sum_{i=1}^n X_i$ et où $\theta = 1.5$, $n = 100$, $k = 90$ et $L = 1000$. La courbe verte présente la loi limite de l'estimateur du maximum de vraisemblance de θ dans le modèle non conditionné.

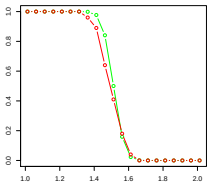


Figure 3: F.d.r. complémentaire en fonction de α

Bibliographie

[1] Bardnoff-Nielsen, O.E. et Cox, R.R. (1990) *Asymptotics Techniques for Use in Statistics*, Chapman and Hall.

[2] Broniatowski M. et Caron, V. (2010) Long runs under point conditioning. The real case. *arXiv:1010.3616,2010.*

[3] Csiszar, I. (1984) Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Prob.* 12 768-793.

[4] Dembo, A. et Zeitouni, O. (1996) Refinements of the Gibbs conditioning principle.*Probab. Theory Related Fields* 104 1-14.

[5] Diaconis, P. et Freedman, D.A. (1988) Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. *J. Theoret. Probab.* 1 381-410.

[6] Lehmann, E. (1986) *Testing Statiscal Hypotheses.* Springer.

Epidémiologie

Méthode de la série de cas : modèles et applications récentes, *Mounia Hocine Hocine, Michel Chavance and Paddy Farrington*

La méthode de la série de cas a été développée par Farrington en 1995 afin d'étudier la liaison entre une exposition intermittente et un événement à survenue aigüe, rare, potentiellement récurrent, en n'utilisant que des sujets cas. Cette méthode a été largement utilisée en pharmaco-épidémiologie, en particulier dans les études de sécurité vaccinale. Elle repose sur un modèle de Poisson conditionné sur le nombre d'événements du sujet et ses antécédents d'exposition. Le conditionnement conduit à un ajustement sur les facteurs constants au cours du temps, et permet ainsi d'éviter certains biais de sélection ou de confusion auxquels sont exposées les enquêtes cas-témoins ou les études de cohorte. Les avantages de la méthode de la série de cas ainsi que ses applications récentes seront présentées.

Waffect : a method to simulate case-control samples in genome-wide association studies, *Vittorio Perduca, Raphaël Mourad, Christine Sinoquet and Gregory Nuel*

Dans une étude d'association à l'échelle du génome, les génomes d'un grand groupe de personnes sont examinés afin d'établir si une association significative existe entre une maladie et les gènes. Le groupe de personnes est divisé en cas (les personnes atteintes) et témoins (ceux qui ne sont pas atteints). L'association est évaluée à travers des tests d'hypothèses. La distribution sous l'hypothèse nulle H0 est étudiée empiriquement en permutant le statut cas/témoins, tandis que pour simuler sous H1 des modèles génétiques complexes sont généralement utilisés. Nous avons développé une approche alternative à ce problème. L'idée est d'imiter les simulations sous H0 en affectant le statut des individus sous la contrainte que la probabilité d'être un cas est compatible avec le modèle de maladie choisi et que le nombre total de cas est fixé. Nous proposons un algorithme simple mais efficace pour effectuer cette affectation. Nous appliquons notre algorithme à des données réelles issues du 1000 genomes project afin de comparer la précision de différentes méthodes pour identifier des marqueurs génétiques causaux. Les résultats montrent que la précision diminue rapidement quand la longueur des régions candidates augmente.

Estimation par imputation multiple du risque relatif et de la capacité prédictive dans les enquêtes cas-cohorte, *Helena Marti and Michel Chavance*

Les estimateurs pondérés utilisés en analyse des études cas-cohorte sont parfois peu efficaces. Or, l'enquête cas-cohorte représente un cas particulier de données incomplètes et des méthodes d'analyse pour données incomplètes peuvent être pertinentes, en particulier, l'imputation multiple. Cette approche est basée sur la génération de plusieurs jeux plausibles de données complètes, prenant en compte l'incertitude sur les données manquantes. Si le modèle d'imputation est correct, l'estimateur de l'imputation multiple est sans biais. Nous avons mon-

tré qu'un modèle d'imputation correct peut être estimé à partir des données complètes (cas et témoins) en utilisant la variable indicatrice des cas parmi les variables explicatives. Nous avons simulé des enquêtes cas-cohorte dont les sous-cohortes étaient sélectionnées par un tirage stratifié. L'imputation multiple et les estimateurs pondérés fournissaient des estimations non-biaisés. Les estimations de l'imputation multiple étaient plus précises que celles obtenues par l'analyse pondérée pour les variables de phase-1, et aussi ou légèrement plus précises pour la variable de phase-2. En outre, l'imputation multiple permet d'estimer, dans les enquêtes cas-cohorte, la capacité prédictive d'un modèle ou celle d'une variable supplémentaire ainsi que la variance de cette estimation.

Initialisation de l'algorithme EM champ-moyen pour les mélanges de Poisson pour données spatiales et application à la cartographie du risque en épidémiologie, *Lamiae Azizi, Florence Forbes, Myriam Charras-Garrido, David Abrial and Senan Doyle*

Dans ce travail nous présentons une nouvelle stratégie d'initialisation pour l'algorithme EM champ-moyen utilisé pour l'estimation des paramètres d'un champ de Markov cache. Nous présentons un modèle de mélange de Poisson pour les données épidémiologiques, ou l'appartenance aux composantes du mélange est modélisé comment étant un modèle de Potts. L'algorithme EM est connu par sa forte dépendance aux conditions initiales. La stratégie d'initialisation proposée est une adaptation, dans un cadre spatial, de celle proposée par Biernacki (2004) pour les modèles de mélanges gaussiens. L'idée est de contraindre les valeurs initiales des paramètres de mélange a être dans l'espace des trajectoires de l'algorithme EM. Nous proposons par la suite une procédure complète d'initialisation permettant d'obtenir des valeurs initiales "raisonnables" des paramètres du champ de Markov. Une étude approfondie sur un ensemble de données simulées et un jeu de données réelles concernant la maladie de l'Encyphalopathie Spongiforme Bovine en France, montrent la performance de cette stratégie comparée aux stratégies d'initialisation classiques.

Gestion des troncatures dans l'analyse des données longitudinales : application à l'étude du développement de la réponse anticorps du paludisme, *Djénéba Thiam, Célia Dechavanne, André Garcia and Grégory Nuel*

Dans ce papier nous présentons une méthode qui combine un algorithme de type Stochastic Expectation Maximization et un algorithme de Gibbs-Sampling pour la gestion des troncatures hautes et basses dans le cadre des modèles linéaires à effets mixtes. Nous comparons à l'aide de simulations cette méthode aux approches d'imputation simple les plus couramment utilisées. La méthode implémentée se montre plus efficace que les méthodes d'imputation simple usuelles, notamment pour l'estimation des effets aléatoires d'un modèle linéaire à effets mixtes.

Ce travail s'intègre dans le cadre de l'étude de cohorte de nouveau-nés menée par l'équipe de l'IRD UMR216 (Santé de la mère et de l'enfant en milieu tropical) au Bénin pour l'analyse de la réponse anticorps de l'enfant entre trois mois et dix-huit mois face au paludisme à Plasmodium

falciparum, avec prise en compte des facteurs environnementaux et cliniques. Cette première étape de l'analyse des données permettra de comprendre en fonction des différentes covariables la variabilité liée à chaque individu de la cohorte.

La méthode de la série de cas : modèles et applications récentes

Mounia Hocine[1], Michel Chavance[2], Paddy Farrington[3]

[1] Conservatoire National des arts et Métiers, 292 rue Saint Martin 75003 Paris

[2] INSERM U1018, CESP, Biostatistique, 16 av. P. V.Couturier 94807 Villejuif

[3] The Open University, Walton Hall, Milton Keynes MK7 6AA, Royaume Uni

La méthode de la série de cas a été développée par Farrington [1] en 1995 afin d'évaluer la liaison entre un événement à survenue aigüe, rare, potentiellement récurrent et une exposition intermittente, en utilisant seulement les données des sujets cas. C'est une alternative aux études épidémiologiques classiques cas-témoins et cohorte. La méthode de la série de cas s'applique lorsque la responsabilité de l'exposition dans la survenue de l'événement d'intérêt n'est plausible que dans un intervalle de temps limité, et s'appuie sur un partage de la période d'observation du sujet en périodes "à risque" au cours desquelles l'exposition est supposée pouvoir entraîner la survenue de l'événement d'intérêt et en périodes "sans risque" où une telle responsabilité est exclue. En pratique, les sujets sont comparés à eux mêmes, comme dans les essais thérapeutiques en cross-over. Concrètement, on suppose que le risque de survenue de l'événement d'intérêt est égal au risque de base durant les périodes sans risque, et qu'il est multiplié par une quantité RR durant les périodes à risque. Ce risque relatif RR est estimé en comparant l'incidence des événements pendant les périodes à risque et les périodes sans risque.

Cet auto-appariement entraîne plusieurs avantages par rapport aux études de cohorte et cas-témoins :

- l’ajustement sur l’ensemble des facteurs de confusion qui ne varient pas au cours du temps (sexe, caractéristiques génétiques, socio-économiques, …) est automatique ;
- la mise en œuvre est facile et rapide puisque aucune information n’est recherchée sur les non cas ;
- les difficultés liées au choix de témoins pertinents sont évitées.

La méthode de la série de cas a été utilisée dans divers domaines de l'épidémiologie, en particulier dans des études de sécurité vaccinale. Elle a également été utilisée en pharmaco-épidémiologie non vaccinale, pour étudier la liaison entre vol de longue durée et thromboembolie veineuse, la relation entre prescriptions de divers médicaments et accidents de circulation, etc.

Des modèles ont été développés récemment pour permettre une plus large utilisation de la méthode de la série de cas lorsque les hypothèses sur lesquelles elle est basée ne sont pas vérifiées. Ces modèles ainsi que plusieurs applications récentes de la méthode seront présentés.

[1] Whitaker HJ, Farrington CP and Musonda P. Tutorial in Biostatistics: The self-controlled case series method. Statistics in Medicine 2006;**25**(10):1768-1797.

Abstract

The case series method was developed by Farrington (1995) to investigate the strength of association between a time-varying exposure and an acute rare event potentially recurrent, using cases only. It can be used when the exposure can only be causally related to the event during a limited period of time and it has been widely used in pharmaco-epidemiology, particularly in the study of vaccine safety. The method is derived from a

Poisson model by conditioning on the individual total number of events and his exposure history. As a consequence of this conditioning the effects of fixed covariates cancel out, and offer to the method a particular advantage as compared to cohort and case-control studies.

Key words: Case series, self-control, acute event, transient exposures, vaccination.

Waffect: a method to simulate case-control samples in genome-wide association studies

Vittorio Perduca[1] & Raphaël Mourad[2] & Christine Sinoquet[3] & Gregory Nuel[1]

[1]*MAP5 - UMR CNRS 8145, Université Paris Descartes, 45 Rue des Saints Pères, 75006 Paris*
vittorio.perduca@gmail.com, gregory.nuel@parisdescartes.fr

[2]*LINA - UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes, 44306 Nantes*
raphael.mourad@univ-nantes.fr

[3]*LINA - UMR CNRS 6241, Université de Nantes, 44322 Nantes*
christine.sinoquet@univ-nantes.fr

Resumé

Dans une étude d'association à l'échelle du génome, les génomes d'un grand groupe de personnes sont examinés afin d'établir si une association significative existe entre une maladie et les gènes. Le groupe de personnes est divisé en cas (les personnes atteintes) et témoins (ceux qui ne sont pas atteints). L'association est évaluée à travers des tests d'hypothèses. La distribution sous l'hypothèse nulle H0 est étudiée empiriquement en permutant le statut cas/témoins, tandis que pour simuler sous H1 des modèles génétiques complexes sont généralement utilisés. Nous avons développé une approche alternative à ce problème. L'idée est d'imiter les simulations sous H0 en affectant le statut des individus sous la contrainte que la probabilité d'être un cas est compatible avec le modèle de maladie choisi et que le nombre total de cas est fixé. Nous proposons un algorithme simple mais efficace pour effectuer cette affectation. Nous appliquons notre algorithme à des données réelles issues du *1000 genomes project* afin de comparer la précision de différentes méthodes pour identifier des marqueurs génétiques causaux. Les résultats montrent que la précision diminue rapidement quand la longueur des régions candidates augmente.

Mots-clés: Échantillonnage forward-backward, phénotype, puissance, aire sous la courbe

Abstract

In a Genome-Wide Association Study (GWAS) the genomes of a large group of individuals are examined to establish the presence of a significant association between a disease and particular genes. The group of individuals is divided in cases (people with the disease) and controls (people without). The association is assessed through statistical hypothesis testing. The distribution under the null hypothesis H0 is empirically studied shuffling uniformly affectation status (case and control memberships), while complex genetic models are usually used to simulate under the alternative hypothesis H1. We have developed an alternative approach to this problem. The idea is to mimic the H0 simulations by affecting status to individuals under the constraint that the probability to be a case is consistent with the chosen disease model and that the total number of cases is fixed. We suggest a simple but efficient algorithm to perform this constrained affectation. We apply our algorithm to a real data set from the *1000 Genomes Project* to compare the accuracy of different methods for identifying causal genetic markers and show that accuracy quickly decreases as the candidate regions get wider.

Keywords: Forward-backward sampling, phenotype, power, area under the curve

1 Introduction

Genome-wide association studies (GWAs) are a widely-used approach to address the localization of causal mutations responsible for common complex genetic diseases, Morris and Cardon (2007). Such studies involve the investigation of hundred of thousands to millions of genetic markers (such as single nucleotide polymorphisms – SNPs), for a cohort of cases and controls whose sizes range in the thousands to tens of thousands individuals. GWASs have met many successes, most notably for type 1 and type 2 diabetes, inflammatory bowel disease, prostate cancer and breast cancer, Hindorff et al. (2009).

In GWASs, very high false positive rates are expected unless a correction for multiple testing is performed. Symmetrically, control for true negative rate - or power - is necessary. Power estimation is the key to evaluate the efficiency of GWAS methods, Spencer et al. (2009). The correct estimation of both rates must take into account the existence of high-dependency patterns between SNPs, or linkage disequilibrium (LD). The accurate estimation of the family wise type I error risk in presence of LD consists in sampling the H_0 distribution through permutations of phenotypes, Zhang and Ott (2010). Thus, any association between loci and the phenotype is broken. This permutation strategy is implemented as a gold standard in numerous toolsets designed for GWASs, e.g. Purcell et al. (2007).

Power is a still more complicated function of several factors: study design, correlation patterns in the genotypic data, sizes of cohorts, frequency of the causal allele, relative risk conferred by the causal factor, genetic model (additive, dominant, recessive, multiplicative), Lettre et al. (2007).

Two main strategies have been developed to simulate H_1: (i) the prospective one, Hyam et al. (2008), which first generates a large sample of haplotypes conditional on reference haplotypes such as HapMap haplotypes, The International HapMap Consortium (2007), then pairs haplotypes to build diplotypes and assigns the disease status depending on the penetrance model involving a randomly selected causal SNP, and (ii) the retrospective strategy, Spencer et al. (2009), which first randomly selects a causal SNP and generates a fixed number of cases and controls, then assigns diplotypes at the causal SNP for cases and controls depending on the penetrance model and finally builds haplotypes (two for each diplotype) for all remaining SNPs of the chromosome, conditionally on reference haplotypes. Nevertheless, both strategies entail implementation problems when applied to power estimation in GWASs. The first strategy presents the drawback to not allow the control of the numbers of cases and controls. To tackle this issue, rejection sampling of case-control samples is used, but leads to a waste of data and time. The second strategy controls these numbers by first fixing them, but requires to build haplotypes for each simulation. The widely-used simulator Hapgen, Su et al. (2010), implements the second strategy.

We propose a new method able to assign exactly n cases and m controls conditionally on the $n + m$ observed genotypes. We first describe the method itself, which we called

WAffect, then illustrate its interest on real GWAS data by comparing the power of several approaches. WAffect will be soon available as an R package. It can affect the phenotypes of 10,000 individuals with 5,000 cases in 2.2 seconds (on a common - even a bit outdated - workstation).

2 Methods

Let $I = \{1, \ldots, q\}$ be the ordered set of all individuals. We denote P_i, $i \in I$, the random variable accounting for the status (phenotype) of individual i, $P_i \in \{0, 1\}$ where 0 stands for *control* and 1 for *case*. Let $\mathbb{P}(P_i = 1) = \pi_i$. We denote N_i, $i \in I$, the random variable counting the total number of cases among individuals indexed by $\{1, \ldots, i\}$: $N_i = \sum_{j=1}^{i} P_j$, with the convention that $N_0 = 0$. Observe that $N_i = N_{i-1} + P_i$ and therefore $N_i \in \{0, \ldots, i\}$ for each i. When all the probabilities π_i are given, we are interested in sampling the values of the P_is, given the condition that the total number of cases N_q must be equal to r, i.e. in sampling the distribution $\mathbb{P}(P_1, \ldots, P_q | N_q = r)$. To achieve this goal we find recursive formulas for the probabilities $\mathbb{P}(P_i = 1 | N_{i-1} = m, N_q = r)$, $i \in I$:

Theorem 1 *For each individual $i = 1, \ldots, q$:*

$$\mathbb{P}(P_i = 1 | N_{i-1} = m, N_q = r) = \frac{\pi_i B_i(m+1)}{B_{i-1}(m)}, \tag{1}$$

where the backward quantities B_i *can be computed using the recursive formulas*

$$B_i(m) = \pi_{i+1} B_{i+1}(m+1) + (1 - \pi_{i+1}) B_{i+1}(m), \tag{2}$$

with the following edge conditions: $B_0(0) = \pi_1 B_1(1) + (1 - \pi_1) B_1(0)$ and $B_q(m) = \delta(m, r)$, δ being the Kronecker's symbol.

The theorem gives a recursive algorithm, which we called WAffect, to sample in the space of all possible configurations of the P_is under the condition that the number of cases must be r and knowing for each individual i his probability π_i to be a case. Simply compute all the backward quantities with Eq. (2), starting from q, and then, starting from the first individual $i = 1$, affect a status for the individual i accordingly to the binomial distribution $P_i \sim \mathcal{B}\left(\frac{\pi_i B_i(N_{i-1}+1)}{B_{i-1}(N_{i-1})}\right)$. The pseudocode is given below. Observe that if $\pi_i = \pi_0$ for each i, then WAffect outputs a permutation of the phenotypes; this is equivalent to simulating under H0. It is possible to simulate affectations in the case of more than two classes by calling recursively WAffect. For instance, in the case of three classes $\{0, 1, 2\}$, start by affecting status 0 versus status $\{1, 2\}$ and then iterate WAffect for the individuals with status $\{1, 2\}$ to affect status 1 versus status 2.

Algorithm 1 WAffect(M, r)

Input: vector of probabilities $\pi = (\pi_i)_{i=1,\dots,q}$, number r of cases
Output: A vector of affectations for the individuals

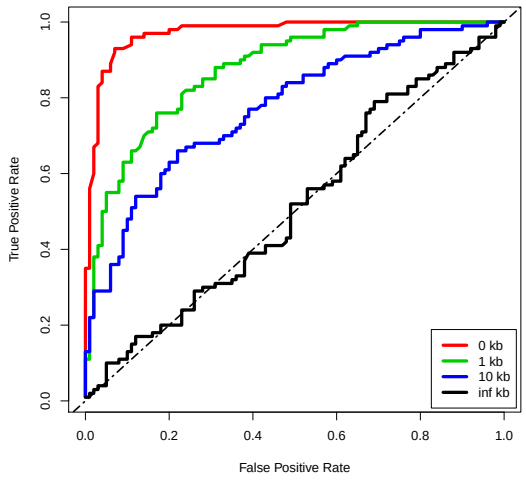
B is $q \times (r+2)$ matrix, $B = 0$ {/* Initialization of B */}
$B_q(r) = 1$
for $i = q-1$ to 0 **do** {/* Iterative computation of B_i */}
  for $m = 0$ to r **do**
    $B_i(m) = \pi_{i+1} B_{i+1}(m+1) + (1-\pi_{i+1}) B_{i+1}(m)$
  end for
end for
Sample P_1 with $P_1 \sim \mathcal{B}\left(\frac{\pi_1 B_1(1)}{\pi_1 B_1(1) + (1-\pi_1) B_1(0)}\right)$ {/* Sampling initialization */}
$N_1 = P_1$
for $i = 2$ to q **do** {/*Sampling of $P_2, \dots, P_q$*/}
  Sample P_i with $P_i \sim \mathcal{B}\left(\frac{\pi_i B_i(N_{i-1}+1)}{B_{i-1}(N_{i-1})}\right)$
  $N_i = N_{i-1} + P_i$
end for
return $P_1, \dots, P_q$

3 Application

We apply WAffect on real data to assess the accuracy of different association methods based on the Cochran-Armitage trend test.

The original data set consists on the real genotypes of 629 individuals from the *1000 Genomes Project*, The 1000 Genomes Project Consortium (2010). We focused on the first 100,000 SNPs on the X chromosome. In the pretreatment stage, we filtered out all the SNPs with Minor Allele Frequency (MAF) less than or equal to 5%, ending up with 8,048 SNPs. Then, we arbitrarily defined an additive disease model with two interacting causal SNPs. In particular, we choose two SNPs S_1, S_2 (having base-pair positions 627,641 and 1,986,325) with low frequencies and showing no linkage disequilibrium (i.e. correlation). For each individual, we defined the following relative risk RR: $RR = 1 + 0.1 \times G_{S_1}$ if $G_{S_2} = 0$, $RR = 1 + 0.1 \times G_{S_2}$ if $G_{S_1} = 0$, $RR = 1.3 + 0.1 \times (G_{S_1} + G_{S_2})$ if $G_{S_1} \times G_{S_2} > 0$, where $G_{S_i} \in \{0, 1, 2\}$ is the number of the less frequent allele in the genotype of S_1, S_2. The last expression defines an *epistasis* (interaction) between the two genes. Then, for each individual we computed his probability π_i to be a case accordingly to his genotype and the disease model. Running 100 times WAffect on the vector $(\pi_i)_{i=1,\dots,629}$ and under the constraint that the total number of cases must be 314, we obtained 100 simulations of the phenotype of each individual under H1. Similarly, we obtained 100 affectations of the phenotypes under H0, running WAffect on a vector whose elements have all the same value in $(0, 1)$ (e.g. $\pi_i = 0.2$ for each i). At last, association analysis was performed

running the toolset PLINK v1.07, Purcell et al. (2007). In particular, for each SNP we obtained the p-value for the trend statistics under H0 and H1. The association methods we studied consider two intervals of varying lengths centered in the two causal SNPs, their statistics being the max of $-\log$ of the p-values in this region. The lengths we considered are ℓ =0, 1, 5, 10, 50 and 100 kb. For each length, we assessed the accuracy of the method analyzing the trade-offs between false positive rate and true positive rate (power). More precisely, for each ℓ we computed the corresponding Receiver Operating Characteristic (ROC) curve, the empirical Area Under the Curve (AUC) value and an upper bounds σ_m for the standard deviation of the estimate, see for instance Metz (1978). The results, depicted in Figure 1, show that we have excellent accuracy when length is 0 kb, and good accuracy up to a length of 1 kb. For greater lengths, AUC $\leq$ 0.8 and accuracy quickly decreases. The association method is therefore typically suitable for testing the association between a given disease and a pair of candidate genes but not for wider regions.



ℓ	AUC	σ_m	Accuracy
0 kb	0.97	0.017	excellent
1 kb	0.87	0.033	good
5 kb	0.79	0.040	fair
10 kb	0.77	0.042	
20 kb	0.75	0.043	
50 kb	0.68	0.047	poor
100 kb	0.62	0.049	fail
∞ kb	0.51	0.050	

Figure 1: ROC curves, AUCs and s.d. upper bounds σ_m at different lengths ℓ

4 Conclusions

We introduced a new iterative algorithm to sample case-control affectations under the constraint that the probability to be a case is consistent with the chosen disease model and that the total number of cases is fixed. This method can be used to simulate under $H1$ for assessing the power of GWAS methods. It generalizes permutations, the gold standard for sampling under H_0. New values of the disease status are generated by permuting them according to weights (the probabilities to be affected). Like permutations, it presents the

advantage that genotypes are fixed and only phenotypes are sampled. Two other possible ways to sample case-control affectations along these lines involve rejection and MCMC algorithms. An R package including all these algorithms will be released shortly. Our algorithm allows to evaluate the performance of GWAS methods using real GWAS data sets. We applied it to data sets from the *1000 Genomes Project* to assess the accuracy of a region candidate approach and showed that, given the chosen design (629 individual) and the modest effect used for the simulation (relative risk of 1.7 in the most favorable case), we have only power to detect the association signal at the gene candidate level. For wider candidate regions, the power drops quickly.

References

[1] Klein R. J. (2007) *Power analysis for genome-wide association studies.* BMC Genetics, 8(58).

[2] Lettre G., Lange C. and Hirschhorn J. N. (2007), *Genetic model testing and statistical power in population- based association studies of quantitative traits*, Genetic Epidemiology, 31(4):358-362.

[3] Metz C. E. (1978) *Basic principles of ROC analysis*, Sem Nuc Med, 8:283–298.

[4] Morris A. P. and Cardon L. R. (2007) *Handbook of statistical genetics*, volume 2, chapter *Whole genome association*, pages 1238-1263, Wiley Interscience, 3rd edition.

[5] Hindorff L. A., Sethupathy P., Junkins H. A., Ramos E. M., Mehta J. P., Collins F. S. and Manolio T. A. (2009) *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*, PNAS, 106(23):9362-9367.

[6] Hyam M. C., Hoggart C., OReilly P., Whittaker J., De Iorio M. and Balding D. (2008), *Fregene: Simulation of realistic sequence-level data in populations and ascertained samples*, BMC Bioinformatics, 9(1):364+.

[7] Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., Bender D., Maller J., Sklar P., de Bakker P. I .W., Daly M. J., Sham P. C. (2007) *PLINK: a toolset for whole-genome association and population-based linkage analysis.* American Journal of Human Genetics, 81. Package PLINK available at http://pngu.mgh.harvard.edu/purcell/plink/

[8] Spencer C. C., Su Z., Donnelly P., and Marchini J. (2009) *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genetics, 5(5):e1000477+.

[9] Su Z., Marchini J., Donnelly P. (2010) *HapGen v2*, http://www.stats.ox.ac.uk/ marchini/software/gwas/hapgen.html

[10] The 1000 Genomes Project Consortium (2010) *A map of human genome variation from population-scale sequencing.* Nature, 467, 1061–1073.

[11] The International HapMap Consortium (2007), *A second generation human haplotype map of over 3.1 million snps*, Nature, 449(7164):851-861.

[12] Zhang Q. and Ott J. (2010) *Handbook on Analyzing Human Genetic Data*, chapter *Multiple comparison/testing issues*, pages 277-287. Springer Berlin Heidelberg.

Estimation par imputation multiple du risque relatif et de la capacité prédictive dans les enquêtes cas-cohorte

Helena Marti & Michel Chavance

CESP, INSERM U1018 Équipe Biostatistique,
16 avenue Paul Vaillant-Couturier, 94807 Villejuif, France
helena.marti-soler@inserm.fr
Tel +33 (0)1 45 59 50 63, Fax +33 (0)1 45 59 51 69

Résumé

Les estimateurs pondérés utilisés en analyse des études cas-cohorte sont parfois peu efficaces. Or, l'enquête cas-cohorte représente un cas particulier de données incomplètes et des méthodes d'analyse pour données incomplètes peuvent être pertinentes, en particulier, l'imputation multiple. Cette approche est basée sur la génération de plusieurs jeux plausibles de données complètes, prenant en compte l'incertitude sur les données manquantes. Si le modèle d'imputation est correct, l'estimateur de l'imputation multiple est sans biais. Nous avons montré qu'un modèle d'imputation correct peut être estimé à partir des données complètes (cas et témoins) en utilisant la variable indicatrice des cas parmi les variables explicatives. Nous avons simulé des enquêtes cas-cohorte dont les sous-cohortes étaient sélectionnées par un tirage stratifié. L'imputation multiple et les estimateurs pondérés fournissaient des estimations non-biaisés. Les estimations de l'imputation multiple étaient plus précises que celles obtenues par l'analyse pondérée pour les variables de phase-1, et aussi ou légèrement plus précises pour la variable de phase-2. En outre, l'imputation multiple permet d'estimer dans les enquêtes cas-cohorte la capacité prédictive d'un modèle ou celle d'une variable supplémentaire ainsi que la variance de cette estimation.

Mots-clés: Enquêtes cas-cohorte, imputation multiple, valeur prédictive.

Abstract

The weighted estimators used for analyzing case-cohort studies are not fully efficient. However, case-cohort studies represent a special type of incomplete data, and methods for analyzing incomplete data could be appropriate, in particular, multiple imputation. This approach is based on the generation of several plausible complete data sets, taking into account the uncertainty about missing values. When the imputation model is correct, it reflects appropriately the distribution of the incomplete variables, respectively among the cases and controls, and the multiple imputation estimator is unbiased. We have shown that a correct imputation model can be estimated from the fully observed data (cases and controls) using the

case status as an explanatory variable. Simulated case-cohort data with subcohort selected by stratified sampling were analyzed. Multiple imputation and weighted estimators provided unbiased estimators. The multiple imputation estimators were slightly more precise than those obtained with weighted analysis, especially when the phase-2 variable was linked to the event occurrence. Moreover, multiple imputation allows estimating the predictive value of a model or of a new variable in case-cohort surveys.

Keywords: Case-cohort designs, multiple imputation, predictive value.

1 Introduction

Les études de cohorte sont de plus en plus souvent utilisées en épidémiologie parce qu'elles sont plus faciles à interpréter en termes de causalité. Généralement les maladies étudiées ont une incidence très faible, et la puissance dépend du nombre de cas. Les études de cas-cohorte permettent d'en réduire le coût au prix d'une perte minime d'efficacité (Langholz 1990).

Les études cas-cohorte sont réalisées en deux phases. 1) La cohorte est sélectionnée par tirage au sort. On recueille l'information de phase-1 sur tous les sujets. Une sous-cohorte est sélectionnée par tirage au sort et la cohorte entière est suivie de manière à identifier la date de survenue du ou des événements d'intérêt. 2) On recueille l'information de phase-2, plus coûteuse, sur tous les cas, qu'ils appartiennent ou non à la sous-cohorte, ainsi que sur les témoins de la sous-cohorte. La sous-cohorte peut être sélectionnée par un tirage uniforme ou stratifié (Prentice 1986, Therneau 1999, Borgan 2000).

La méthode usuelle d'analyse des enquêtes cas-cohorte est l'analyse pondérée, décrite initialement par Prentice (1986). Or, l'enquête cas-cohorte peut aussi être vue comme un cas particulier de données manquantes au hasard puisque le processus d'observation est contrôlé par les organisateurs de l'étude. Ainsi, des méthodes d'analyse pour données incomplètes peuvent être pertinentes, en particulier, l'imputation multiple.

En outre, l'estimation de la valeur prédictive d'un modèle ou d'une nouvelle variable dans le cadre des enquêtes cas-cohorte est peu développée. Les mesures classiques proposées dans les études de cohorte ne sont pas adaptées à ce type d'échantillonnage. Cependant, tous les outils proposés dans le cadre des études de cohorte peuvent être appliqués aux enquêtes cas-cohorte grâce à l'imputation multiple.

L'objectif de ce travail était de mettre en œuvre l'imputation multiple pour analyser les enquêtes cas-cohorte et pour estimer la valeur prédictive d'un modèle ou d'une variable supplémentaire. Nous avons validé l'approche par des simulations et nous présentons les résultats de l'analyse d'une étude cas-cohorte issue de la cohorte 3C.

2 Observations incomplètes et imputation multiple

Little et Rubin (1987) proposent de distinguer trois processus d'observation: données manquant complètement aléatoirement (MCA) lorsque la probabilité qu'une observation soit incomplète est constante; données manquant aléatoirement (MA) lorsque cette probabilité ne dépend que de valeurs observées; et données manquant non aléatoirement (MNA) lorsque cette probabilité dépend de valeurs non observées.

Les observations incomplètes posent des problèmes de biais, de précision et de puissance. Si on limite l'analyse aux seules observations complètes on s'expose à un biais de sélection pour les processus d'observation MA et MNA. Avec des données MA et une méthode d'analyse pertinente, il est possible d'effectuer des inférences correctes. Le plan cas-cohorte est l'une des rares situations où l'on peut affirmer que les données sont MA car l'observation ne dépend que du statut cas-témoin (échantillonnage uniforme) et éventuellement de variables observées (échantillonnage stratifié).

L'imputation multiple permet d'obtenir une approximation de l'estimateur du maximum de vraisemblance partielle. Sous l'hypothèse de données MA, elle permet de : a) Corriger le biais, b) Obtenir une variance asymptotique correcte. Cette méthode repose sur la génération de plusieurs (M) jeux plausibles de données complètes, en prenant en compte tous les niveaux d'incertitude concernant les valeurs manquantes. On ne remplace pas les données manquantes par leur espérance mais par une valeur tirée dans la loi postulée par le modèle, et, pour tenir compte de l'incertitude sur les paramètres du modèle d'imputation, on effectue plusieurs imputations avec des paramètres tirés dans la loi asymptotique de l'estimateur obtenu à partir des observations complètes.

On obtient une estimation du paramètre d'intérêt $\hat{\theta}_m, \quad m = \{1, \ldots, M\}$ et une estimation de la variance de l'estimateur, $\widehat{V}(\hat{\theta}_m)$ pour chaque jeu de données complétées. Si le modèle d'imputation est correct, les estimateurs $\hat{\theta}_m$ sont non-biaisés. Puis, on obtient une estimation unique moyenne de ces M estimateurs, qui est aussi non-biaisée:

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$$

On peut, grâce à la multiplicité des imputations, estimer correctement la variance de cet estimateur unique, formée par deux composantes: La composante *intra-imputations* (W_{IM}) et la composante *inter-imputations* (B_{IM}):

$$\widehat{V}(\hat{\theta}_{IM}) = \widehat{W}_{IM} + \widehat{B}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \widehat{V}(\hat{\theta}_m) + (1 + M^{-1}) \frac{\sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{IM})(\hat{\theta}_m - \hat{\theta}_{IM})'}{M - 1}$$

L'imputation multiple demande de modéliser correctement les relations entre les variables incomplètes et les variables qui lui sont liées. Nous avons montré que le modèle

d'imputation peut être estimé sur tous les sujets, cas et non-cas, en introduisant une indicatrice des cas comme variable explicative. En pratique, en plus de l'indicatrice de cas et des variables de stratification, il est nécessaire d'ajuster sur les variables explicatives du modèle de Cox et sur d'autres variables prédictives, si elles sont disponibles (Marti et Chavance, 2011).

3 Capacité prédictive d'un modèle

L'évaluation de la capacité prédictive pose des problèmes spécifiques lorsque la variable réponse est censurée. Nous nous sommes intéressés à l'indice C de Harrell (1982), qui est une adaptation de la statistique U de Mann-Withney en présence de censure. La statistique U mesure la concordance entre deux variables X et Y. Elle est fondée sur la comparaison de toutes les paires de sujets avec ($X_i > X_j$ et $Y_i > Y_j$ ou $X_i < X_j$ et $Y_i < Y_j$), où i et j indicent les observations. Néanmoins, pour les données de survie, il n'est pas possible de comparer toutes les paires de sujets à cause du processus de censure.

Soit T_i le délai de survie pour un sujet i, $i = 1, \ldots, n$ et C_i le délai de censure pour un sujet i. On observe $X_i = min(T_i, C_i)$. Soit D_i le délai de survie prédit, et Y_i la probabilité de survie prédite au temps t (durée de l'étude). Les paires formées par deux sujets censurés n'apportent pas d'information sur leur concordance avec la survie prédite par le modèle car les délais de survie sont inconnus. De même, les paires formées par un sujet dont la date de décès est connue et un sujet censuré avant le décès, n'apporte pas d'information sur la concordance car le délai de survie inconnu peut être inférieur ou supérieur au délai de survie observé. Harrell a donc défini comme paires utilisables celles dont les délais de survie prédits et observés sont comparables. L'indice de Harrell est la proportion de paires concordantes parmi les paires utilisables:

$$C = \frac{\pi_c}{\pi_c + \pi_d} \tag{1}$$

où π_c est la probabilité de concordance de la paire (i, j), $i \neq j$, $i, j = 1, ..., n$:

$$\pi_c = P(X_i < X_j \ \texttt{et} \quad Y_i < Y_j) + P(X_i > X_j \ \texttt{et} \quad Y_i > Y_j) \tag{2}$$

Nous supposons que le délai de survie et la probabilité de survie prédite sont continus, d'où $P(X_i = X_j) = P(Y_i = Y_j) = 0$. Alors, de façon analogue, la probabilité de discordance π_d est:

$$\begin{aligned} \pi_d &= P(X_i < X_j \ \texttt{et} \quad Y_i > Y_j) + P(X_i > X_j \ \texttt{et} \quad Y_i < Y_j) \\ &= 1 - \pi_c \end{aligned} \tag{3}$$

Nous avons montré que dans le cadre des enquêtes cas-cohorte, l'application naive de l'indice C n'est pas adéquate car les paires utilisables ne peuvent être définies que pour

les sujets appartenant à l'échantillon cas-cohorte, ce qui peut entraîner des estimations biaisées de la capacité prédictive du modèle. Cependant, nous avons aussi montré, à l'aide des simulations, que l'imputation multiple permet d'estimer la capacité prédictive d'un modèle et d'une nouvelle variable ainsi que leurs variances dans le cadre des enquêtes cas-cohorte.

4 Résultats

Nous avons validé l'imputation multiple par des simulations et nous présentons l'application dans une étude cas-cohorte issue de la cohorte 3C. Brièvement, des sujets âgés de 65 ans et plus ont été recrutés entre 1999 et 2001 dans trois villes françaises: Bordeaux, Dijon et Montpellier. La taille de la cohorte était de $N = 9.294$. Une étude cas-cohorte a été menée visant à évaluer la relation entre le niveau plasmatique de D-dimère (un marqueur de coagulation et fibrinolyse), et les risques d'événement coronarien (EC) et de démence vasculaire (DV) à 4 ans. La cohorte à été stratifiée sur le sexe, la ville et l'âge. L'information de phase-1 concernait les caractéristiques socio-démographiques, le niveau d'éducation, la consommation d'alcool et de tabac. La tension artérielle, le poids et la taille étaient aussi recueillis. Une sous-cohorte de taille $n = 1.254$, (13,5% de la cohorte complète) a été échantillonnée par tirage stratifié sur le sexe, la ville et l'âge. L'incidence d'EC et DV était respectivement de 2% et 0,6%. Les niveaux plasmatiques de D-dimère étaient disponibles seulement pour les sujets de phase-2. Carcaillon (2009) a observé dans cette enquête une augmentation linéaire du risque de DV par rapport au quintiles des niveaux de D-dimère.

Nous avons réanalysé ces données par imputation multiple et par estimateurs pondérés, puis nous avons évalué la capacité prédictive du D-dimère sur le risque de DV. À cause du petit nombre d'événements, nous n'avons pas utilisé les quintiles de D-dimère mais les tertiles. Nous avons construit un modèle d'imputation prenant en compte la relation entre les niveaux plasmatiques de D-dimère et les variables de phase-1: l'indicatrice de cas, les variables de stratification, le niveau éducationnel, l'IMC et l'apolipoprotéine $\epsilon 4$ (apoe $\epsilon 4$). Les estimations fournies par l'imputation multiple étaient légèrement inférieures aux estimations fournies par les estimateurs pondérés et légèrement plus précises. Le risque de DV augmentait significativement avec les niveaux de D-dimère (p de tendance = 0,016).

Nous avons estimé la capacité prédictive de la variable de phase-2, D-dimère. C_1 était la valeur prédictive du modèle de Cox M_1, ajusté sur les variables de phase-1 âge, sexe, centre, niveau éducationnel, et apolipoprotéine $\epsilon 4$ (apoe $\epsilon 4$). C_2 était la valeur prédictive correspondant au modèle M_1 plus la variable additionnelle D-dimère. La valeur prédictive du D-dimère était Δ, la différence entre C_2 et C_1. La valeur prédictive du modèle incluant seulement les variables de phase-1 étaient 0,86. L'addition du D-dimère n'améliorait pas significativement la valeur prédictive du modèle, $\Delta = 0{,}009$, 95%CI = (-0,011; 0,029).

5 Conclusion

En conclusion, l'imputation multiple, menée prudemment, est une alternative simple à l'approche pondérée. Elle permet d'estimer correctement les risques relatifs ainsi que leurs erreurs standard. Les estimations sont plus précises pour les variables de phase-1 et autant ou légèrement plus précises pour la variable de phase-2. Par ailleurs, l'imputation multiple permet d'estimer dans les enquêtes cas-cohorte, la capacité prédictive d'un modèle ou d'une variable supplémentaire ainsi que leurs variances.

Bibliographie

[1] Langholz, B. et Thomas, D. C. (1990), Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison, Am J Epidemiol, 131, 169–176.
[2] Prentice, R.L. (1986), A case-cohort design for epidemiologic cohort studies and disease prevention trials, Biometrika, 73, 1–11.
[3] Therneau, T. M. et Li, H. (1999), Computing the Cox model for case cohort designs, *Lifetime Data Anal*, 5, 99–112.
[4] Borgan, O. *et al.* (2000), Exposure stratified case-cohort designs, *Lifetime Data Anal*, 6, 39–58.
[5] Little, R.J.A. et Rubin, D.B. (1987), *Statistical analysis with missing data*, Wiley, New York.
[6] Marti, H et Chavance, M. (2011), Multiple imputation analysis of case–cohort studies, *Statistics in Medicine* (sous presse).
[7] Harrell, F.E. *et al.* (1982), Evaluating the yield of medical tests, *JAMA*, 247, 2543-2546.
[8] Carcaillon, L. *et al.* (2009), Elevated plasma fibrin D-dimer as a risk factor for vascular dementia: the Three-City cohort study, *Journal of thrombosis and haemostasis*, 7, 1972–1978.

Initialisation de l'algorithme EM champ-moyen pour les Mélanges de Poisson pour données spatiales et application à la Cartographie du risque en Épidémiologie.

Lamiae Azizi[1,2], Florence Forbes[1], Doyle Senan[1], Myriam Charras-Garrido[2] & David Abrial[2]

[1] *INRIA Rhône-Alpes & Laboratoire Jean Kuntzmann, Equipe Mistis, Inovallée, 655 av. de l'Europe, Montbonnot, 38334 Saint-Ismier Cedex.*

[2] *INRA, Unité dÉpidémiologie Animale, Centre de recherche de Clermont-Ferrand-Theix, 63122 Saint-Genès Champanelle.*

Résumé

Dans ce travail nous présentons une nouvelle stratégie d'initialisation pour l'algorithme EM champ moyen utilisé pour l'estimation des paramètres d'un champ de Markov caché. Nous présentons un modèle de mélange de Poisson pour les données épidémiologiques, oú l'appartenance aux composantes du mélange est modélisé comment étant un modèle de Potts. L'algorithme EM est connu par sa forte dépendance aux conditions initiales. La stratégie d'initialisation proposée est une adaptation, dans un cadre spatial, de celle proposée par Biernacki (2004) pour les modèles de mélanges gaussiens. L'idée est de contraindre les valeurs initiales des paramètres de mélange à être dans l'espace des trajectoires de l'algorithme EM. Nous proposons par la suite une procédure compléte d'initialisation permettant d'obtenir des valeurs initiales "raisonnables" des paramètres du champ de Markov. Une étude approfondie sur un ensemble de données simulées et un jeu de données réelles concernant la maladie de l'Encyphalopathie Spongiforme Bovine en France, montrent la performance de cette stratégie comparée aux stratégies d'initialisation classiques.

Mots clés : Champs de Markov cachés, Le variationnel EM, Mélange de Poisson, Cartographie du risque.

Abstract

In this work, we are interested in presenting a new strategy of initialisation for the variational EM mean-field algorithm used for the estimation of a Hidden Markov model parameters. We present a model based on finite mixture Poisson, in which the allocation to the mixture components is modelled through a spatially correlated process, the Potts model. The EM algorithm is suffering from its high dependence to the starting values. The proposed strategy is an extension in a spatial context, to the one proposed by Biernacki (2004) in the multivariate Gaussian mixture context. It consists on randomly drawing initial mixture parameters in an appropriate space including all possible EM trajectories. We complete this search procedure by proposing a strategy which permits us to obtain a reasonable starting values for the

Markov field parameters. An intensive study on a set of simulated and real data shows that this strategy outperforms the classical methods.
Keywords : Hidden Markov field, Variationnal EM, Poisson mixture, Disease mapping.

Introduction

L'objectif principal de la cartographie du risque en épidémiologie est de restaurer la réalité sous-jacente à partir des observations dont on dispose plus une estimation raisonnable du risque en chaque point de la zone étudiée. On suppose que nos données observées Y proviennent d'une population hétérogène et que l'affectation d'une observation à une classe est inconnue. Les modèles de mélanges semblent être un bon candidat pour résoudre ce genre de problème et considèrent que la population sous-jacente est un ensemble de composantes avec différents niveaux de risque. Les données dont on dispose en épidémiologie sont des données de comptage qui seront modélisées par un modèle de mélange de Poisson. L'image à restaurer définit un champ $Z = Z_i, i \in S$ qui sera modélisée dans notre travail par un modèle de Potts à K couleurs. L'estimation des paramètres du modèle est faite par une approche variationnelle de l'algorithme Expectation-maximisation (EM) basée sur la théorie du champ moyen. Cet algorithme est encore plus sensible à l'initialisation dans le cas des mélanges de Poisson que dans un cadre gaussien. Les méthodes d'initialisation usuelles ne donnent pas des résultats satisfaisants dans la plupart des cas et spécialement dans le cas des maladies animales ou le taux de risque est relativement petit. Nous proposons une adaptation de la méthode d'initialisation proposé par Biernacki (2004) pour des mélanges gaussiens à notre cas d'étude.

1 Modèle de Cartographie du risque à l'aide des champs de Markov cachés

Les données dont on dispose en épidémiologie sont respectivement le nombre de cas infectés (malades ou morts..) $y_i (i = 1, \ldots, n)$ pour une maladie donnée et l'effectif de la population cible (données démographiques) n_i pour chaque unité géographique. L'hypothèse clé est que le nombre de cas y_i suit une loi de Poisson de paramètre λ_i avec :

$$f(Y_i = y_i|\lambda_i) = \exp(n_i\lambda_i)\frac{(n_i\lambda_i)^{y_i}}{y_i!}. \quad (1)$$

L'objectif de la cartographie est d'affecter chaque unité i à un des K différents niveaux de risque qui sont inconnus et ont besoin d'être estimés. En général, les niveaux de risque sont supposés être semblables pour les régions voisines. L'idée est d'exploiter cette information spatiale pour une meilleure estimation du risque. En

prenant en compte ces informations, les données sont divisées en données observées $y = \{y_1, \ldots, y_n\}$ et en données non observées (ou manquantes) $z = \{z_1, \ldots, z_n\}$. La dépendance spatiale entre les variables aléatoires voisines Z_i est prise en compte en assumant que la distribution jointe de $\{Z_1, \ldots, Z_n\}$ est un modèle de Potts à K couleurs défini sur un graphe prédéfini G avec :

$$P_G(z|\alpha, \beta) = W(\alpha, \beta)^{-1} \exp(-H(z|\alpha, \beta))$$

; avec W est la constante de normalisation, (α, β) sont les paramètres du champ de Markov caché et :

$$H(z|\alpha, \beta) = \sum_{i \in S} V_i(z_i) + \sum_{i \sim j} V_{ij}(z_i, z_j)$$

.

Dans notre modèle, $V_i(z_i) = -\alpha_{z_i}$ et $V_{ij}(z_i, z_j) = \begin{cases} 1 & \text{si } z_i = z_j \\ \frac{1}{2} & \text{si } |z_i - z_j| = 1 \\ 0 & \text{sinon.} \end{cases}$

2 L'algorithme EM champ moyen pour champs de Markov cachés

En cartographie de risque, le but est de restaurer l'image z interprétée comme une classification du risque en K classes. Les étiquettes sont inconnues et considérées comme des données manquantes. L'objectif est de classer chaque région dans une des K classes. Pour cela, on considère la stratégie de la **maximisation des probabilités marginales** (MPM) qui consiste à maximiser $\forall i, P(Z_i = z_i|y)$. L'algorithme EM est le plus approprié pour l'estimation des paramètres en présence des données manquantes. Dans le cas des champs de Markov cachés, et à cause des dépendances spatiales, l'algorithme EM a besoin d'approximations. On propose d'utiliser des approximations basées sur la théorie du champ moyen Celeux (2003). L'idée est de se ramener à un système de variables indépendantes en négligeant les fluctuations des voisins d'un site i : En résumé faire comme si ces voisins prennent leur valeur moyenne. L'algorithme se décompose à l'itération q en 2 étapes :

- Générer a partir des données y et des estimations courantes des paramètres $\theta^{(q-1)} = (\alpha^{(q-1)}, \beta^{(q-1)}, \lambda^{(q-1)})$, une configuration $\tilde{z}^{(q)} = (\tilde{z}_1^{(q)}, \ldots, \tilde{z}_n^{(q)})$ pour les Z_i et approximer la distribution $P_G(z|\alpha, \beta)$ par $P_{\tilde{z}^{(q)}}(z|\alpha, \beta) = \prod_{i \in S} P_G(z_i|\tilde{z}^{(q)}_{\mathcal{N}(i)}; \alpha, \beta)$.
- Appliquer l'algorithme EM sur le modèle factorisé $P_G(y, z|\theta) \approx \prod_{i \in S} f_i(y_i|z_i, \lambda) P_G(z_i|\tilde{z}^{(q)}_{\mathcal{N}(i)}; \alpha, \beta)$ à partir des estimations $\theta^{(q-1)}$ en vue d'obtenir de nouvelles estimations $\theta^{(q)}$:

  1. **étape E** Calcul des probabilités *a posteriori* pour tout i et k :

$$\tilde{t}_{ik}^{(q)} = P(Z_i = k|\mathbf{y}, \tilde{z}^{(q)}_{\mathcal{N}(i)}; \theta^{(q-1)}).$$

2. **étape M** Mise à jour des paramètres θ :

$$\text{for all k,} \quad \lambda_k^{(q)} = \frac{\sum_{i \in S} \tilde{t}_{ik}^{(q)} \, y_i}{\sum_{i \in S} n_i \, y_i} , \tag{2}$$

and

$$(\alpha, \beta)^{(q)} = \arg\max_{\alpha,\beta} \sum_{i \in S} \sum_{k=1}^{K} \tilde{t}_{ik}^{(q)} \log \tilde{\pi}_{ik}^{(q)}, \tag{3}$$

ou $\tilde{\pi}_{ik}^{(q)} = P(Z_i = k | \tilde{z}_{\mathcal{N}(i)}^{(q)}; \alpha, \beta)$.

3 Initialisation de L'algorithme EM champ moyen

L'algorithme EM souffre de sa grande sensibilité aux conditions initiales, encore plus quand il s'agit d'un modèle de mélange de poisson avec des moyennes petites. Plusieurs stratégies d'initialisation ont été proposées dans le cas des mélanges gaussiens pour s'affranchir de cette limite. Dans ce travail, nous proposons une adaptation de la stratégie d'initialisation prenant en compte les trajectoires de l'algoritme EM proposé par Biernacki pour les modèles de mélanges gaussiens. L'idée de cette stratégie est de s'assurer que les valeurs initiales tirées explorent largement l'espace des trajectoires de l'algorithme EM et s'approchent le plus possible des vrais valeurs.
Soit $n = \sum_{i \in S} n_i$ l'effectif de la population totale. A chaque itération q, on a

$$n_k^{(q)} = \frac{\sum_{i \in S} \tilde{t}_{ik}^{(q)} n_i}{n},$$

qui est interprété comme la proportion de la population qui appartient au niveau de risque k. Il s'en suit que $\sum_{k=1}^{K} n_k^{(q)} = 1$ et :

$$\sum_{k=1}^{K} n_k^{(q)} \lambda_k^{(q)} = \frac{\sum_{i \in S} y_i}{n} = \bar{\lambda}.$$

$\bar{\lambda}$ peut être interprété comme un risque moyen et a l'avantage de ne dépendre que des données observées. A chaque itération q, l'estimation courante du paramètre $\lambda_k^{(q)}$ satisfait cette équation et tous les trajectoires de l'algorithme EM sont supposées être dans l'espace défini par cette équation. Donc l'idée serait d'avoir des valeurs initiales de λ_k tirées dans cette espace. la procédure de tirage se fait en 2 étapes :

1. On tire des valeurs de $n_k^{(0)}$ selon une distribution de Dirichlet $\mathcal{D}(\pi, \ldots, \pi)$ avec $\pi = 1$ pour une distribution uniforme dans l'espace défini par $\sum_{k=1}^{K} n_k^{(0)} = 1$.
2. On tire uniformément et sans répétition des valeurs de $\lambda_l^{(0)}$ dans l'échantillon $\{\frac{y_1}{n_1}, \ldots, \frac{y_N}{n_N}\}$ sauf pour une des composantes k du mélange (choisie par l'utilisateur) qui vérifie l'équation :

$$\lambda_k^{(0)} = \frac{\bar{\lambda} - \sum_{l \neq k} n_k^{(0)} \lambda_l^{(0)}}{n_k^{(0)}}.$$

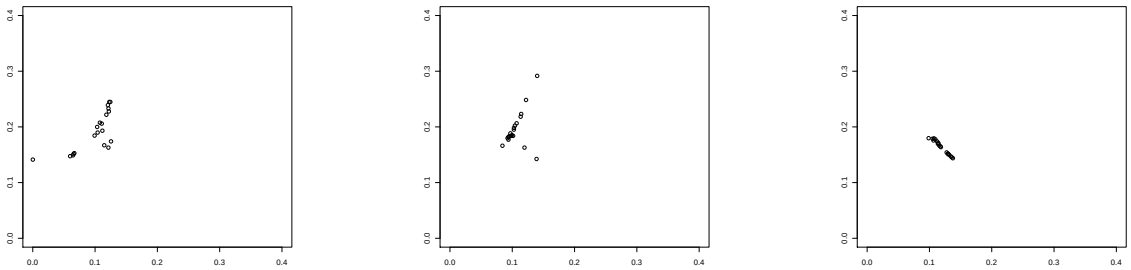
Ces étapes génèrent un vecteur de paramètres. Le nombre de valeurs initiales obtenu est décidé par l'utilisateur en fonction de son usage. Cette procédure de recherche ne garantit en rien la positivité des valeurs générées. Si ce n'est pas le cas, la valeur simulée est écartée et on réitère la procédure.
Pour illustrer cette stratégie, on la compare à 2 autres stratégies qui sont l'initialisation aléatoire et l'initialisation avec une partition déterminée au préalable. Pour cela, on prend un exemple de mélange indépendant avec 2 classes de moyennes $\lambda_1 = 0.1$ et $\lambda_2 = 0.2$ et de proportions $\pi_1 = \pi_2 = 0.5$. Les valeurs y_i sont simulées selon le modèle (1) et avec des n_i tirés aléatoirement et sans répétition entre 10 et 109. On génère 100 valeurs initiales avec les 3 stratégies comparées. Figure 1 montre les valeurs obtenues après 1 itération de l'algorithme EM de 20 valeurs choisies (plus facile à visualiser) parmi les 100 générées par les 3 différentes stratégies. On peut voir qu'avec la stratégie proposée, on obtient des valeurs proches des vraies alors qu'avec l'initialisation aléatoire, les valeurs sont plus dispersées et avec l'initialisation par partition les valeurs sont regroupées n'importe où. Cette stratégie ne nous fournit que des valeurs initiales des paramètres λ et il nous reste à initialiser "raisonnablement" les paramètres α et β. Pour ce faire, nous proposons une procédure complète de recherche de raisonnables valeurs initiales des paramètres θ, qui se décompose en 2 étapes :

1. Générer un nombre M de valeurs initiales λ^0 selon la procédure de tirage précédente.
2. Pour chaque valeur initiale générée (λ^0), on pose $\alpha^0 = 0$ et $\beta^0 = 0$. On tourne l'algorithme EM champ moyen jusquâà sa convergence en gardant la valeur de β fixée à sa valeur initiale 1.

A l'issue de ces 2 étapes, on obtient un ensemble de valeurs initiales pour l'ensemble des paramètres θ. On fait tourner notre algorithme EM champ moyen en initialisant avec cet ensemble de paramètres obtenu jusquâà sa convergence. On retient au final le résultat associe à la plus grande vraisemblance.

4 Illustration numérique

L'objectif de notre étude est de traiter les maladies animales rares pour lesquelles le nombre de cas observé est moindre que 10 par unité géographique. Pour illustrer



(a) (b) (c)

FIGURE 1 – 20 valeurs de (λ_1, λ_2) obtenus après 1 itération de l'algorithme EM quand, (a) : les valeurs initiales ont été générées aléatoirement ; (b) : les valeurs initiales ont été générées avec notre stratégie ; (c) : les valeurs initiales ont été générées selon une partition.

la performance de notre modèle, nous présenterons les résultats obtenus sur des données simulées et des données réelles. Pour les données réelles nous étudions l'Encyphalopatie Spongiforme Bovine en France entre 2001 et 2005. Dans les résultats présentés pour les données simulées, nous comparons la stratégie d'initialisation proposée avec les stratégies usuelles et montrons l'efficacité de celle- là dans la majorité des situations susceptibles de se produire en réalité. Malgré la difficulté des données étudiées et en présence d'un nombre important de 0, notre modèle initialisé par la stratégie proposée donne des résultats satisfaisants.

Bibliographie

[1] Biernacki, C. Celeux, G. et Govaert, G. (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics Data Analysis*, 41, 561–575.
[2] Biernacki, C. (2004) Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures. *Statistics and Computing*, 14 :3, 267-279.
[3] Celeux, G. Forbes, F. et Peyrard, N. (2003) EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36, 131-144.

Gestion des troncatures dans l'analyse des données longitudinales : application à l'étude du développement de la réponse anticorps du paludisme.

Djénéba Thiam*[1] , Célia Dechavanne[2] , André Garcia[2] , Grégory Nuel[1]

[1]*Laboratoire MAP5 , 45 rue des Saints Pères, 75270 Paris Cedex 6, France*
djeneba.thiam-diarra@parisdescartes.fr

[2]*Institut de Recherche pour le Développement IRD / UMR216, Laboratoire de parasitologie, 4 avenue de l'observatoire, 75270 Paris Cedex 6, France*

Abstract

In this paper we introduce a methodological approach to cover the problem of truncation in the response variable of a linear mixed model. We show how to overcome this problem using a Stochastic Expectation Maximization (SEM) algorithm associated with Gibbs-sampling techniques. We compare our new approach to several popular heuristics for this problem (remove truncated data, replace truncated data by the detection level or by half the detection level) and show how the SEM supersedes these methods.

This work is closely connected to an research team in epidemiology UMR216 (Mother's and Child's Health in Tropical Environments, IRD (*Institut de Recherche pour le Développement*). The aim of the UMR216's study in Tori-Bossito in Bénin is to understand the newborn antibody response against malaria. We validate our methodology using these data.

Résumé

Dans ce papier nous présentons une méthode qui combine un algorithme de type *Stochastic Expectation Maximization* et un algorithme de *Gibbs-Sampling* pour la gestion des troncatures hautes et basses dans le cadre des modèles linéaires à effets mixtes. Nous comparons à l'aide de simulations cette méthode aux approches d'imputation simple les plus couramment utilisées. La méthode implémentée se montre plus efficace que les méthodes d'imputation simple usuelles, notamment pour l'estimation des effets aléatoires d'un modèle linéaire à effets mixtes.

Ce travail s'intègre dans le cadre de l'étude de cohorte de nouveau-nés menée par l'équipe de l'IRD UMR216 (Santé de la mère et de l'enfant en milieu tropical) au Bénin pour l'analyse de la réponse anticorps de l'enfant entre trois mois et dix-huit mois face au paludisme à *Plasmodium falciparum*, avec prise en compte des facteurs environnementaux et cliniques. Cette première étape de l'analyse des données permettra de comprendre en fonction des différentes covariables la variabilité liée à chaque individu de la cohorte.

Mots-clefs : Troncature ; algorithme SEM ; *Gibbs-Sampling* ; modèles linéaires à effets mixtes.

Introduction

Le problème de données non observées en raison d'une limite de la technique utilisée est courant dans l'étude des données longitudinales. En biologie les valeurs non observées en dessous du seuil de détection de l'outil de mesure sont le plus souvent supprimées ou remplacées par le seuil de détection lui même ou la moitié du seuil de détection. Ces méthodes sont connues pour introduire du biais au niveau des estimateurs et entraîner perte d'information. Le biais induit est d'autant plus important que le taux de troncature est élevé.

Maximiser la vraisemblance en présence de données incomplètes est difficile voire impossible à faire analytiquement. Aussi a t'on recours à des algorithmes numériques d'optimisation ou à des méthodes d'imputation multiples. Diverses approches ont été traitées dans la littérature pour y faire face dans le cadre des modèles linéaires à effets mixtes. Nous pouvons citer les algorithmes de type EM (Hughes, 1999). Il y a aussi les approches plus directes qui font appel à des algorithmes d'optimisation directe de la vraisemblance (Jaqmin-Gadda and al, 2000). L'efficacité de ces méthodes dépend du nombre d'observations, du nombre de covariables, du nombre d'effets aléatoires (Thiebaut and Jaqmin-Gadda, 2004).

Le problème de données non observées a aussi été étudié dans le cadre des modèles non linéaires à effets mixtes. La méthodologie est proche de celle développée par Hughes (1999) dans le cas linéaire et consiste à utiliser des extensions de l'algorithme EM pour gérer ce problème (Samson and al, 2006).

Nous nous proposons de mettre en œuvre un algorithme de type *Stochastic Expectation Maximization* (SEM) associé à un algorithme de type *Monte Carlo Chain of Markov* (MCMC), le *Gibbs-Sampling*.

Nous appliquons ensuite la méthode implémentée aux données de la cohorte de Tori-Bossito sur l'étude de la réponse anticorps du nouveau-né face au paludisme, données sur lesquelles on observe environ 30% de troncature.

Méthodes

Algorithme *Stochastic Expectation Maximization*

Les mesures complètes Y d'un individu i sont composées de mesures non tronquées X et de mesures tronquées S : $Y = (X, S)$.

Algorithme1 SEM : N nombre d'itérations, M valeur de *"burn-in"* [1].

1) Initialisation arbitraire de $\Theta^{(0)}$
2) Pour $j \in \{1, \ , \ldots, \ N\}$
3) Tirage aléatoire de $S = s^j \sim f_{\Theta^{(j)}}(S|X)$
4) $\Theta^{(j+1)} = \arg\max_{\Theta} \ell(\Theta|X = x, S = s^j)$
5) Fin for
6) $\widehat{\Theta} = \frac{1}{N-M} \sum_{N}^{j=M} \Theta^{(j)}$

[1] Il se passe un temps de chauffe avant que la chaîne de Markov entre dans sa phase stationnaire : c'est la phase transitoire de la chaîne ; durant cette phase les paramètres se cherchent ces valeurs sont dites valeurs de burn-in

La loi $f(S|X)$ pouvant parfois être très compliquée, on doit parfois recourir à des approximations numériques comme le *Gibbs-Sampling* pour simplement être capable d'échantillonner sous cette loi. Ainsi la partie 3 de l'algorithme SEM est réalisée en utilisant la méthode de *Gibbs-Sampling*, qui permet de tirer les valeurs non observées conditionnellement aux valeurs observées et en respectant la contrainte de seuil de détection ou de saturation.

Le processus $(\Theta^j)_{j\geq 0}$ est un processus de Markov d'ordre 1 dont l'espérance sous la loi stationnaire est l'estimateur du maximum de vraisemblance (Celeux and Diebolt, 1988).

Algorithme du *Gibbs-Sampling*

Considérons l'exemple de trois variables aléatoires .

Soit $X = (X, Y, Z)$, le *Gibbs-Sampling* (Casella and George, 1992) génère un échantillon de loi $f(x)$ sans passer par le calcul de la densité, en passant par les lois conditionnelles $f(X|Y, Z), f(Y|X, Z), f(Z|Y, X)$. Ainsi nous obtenons une séquence de variables aléatoires :

$$X^{(0)}, Y^{(0)}, Z^{(0)}, \ldots, X^{(k)}Y^{(k)}, Z^{(k)} \tag{1}$$

$X^{(0)}$, $Y^{(0)}$,$Z^{(0)}$ sont initialisés par une valeur arbitraire et ainsi à chaque itération les variables aléatoires $X^{(j)}$, $Y^{(j)}$, $Z^{(j)}$ sont tirées à partir des lois conditionnelles comme suit :

$$\begin{aligned} X^{(j+1)} &\sim f\left(X|Y^{(j)}, Z^{(j)}\right) \\ Y^{(j+1)} &\sim f\left(Y|X^{(j+1)}, Z^{(j)}\right) \\ Z^{(j+1)} &\sim f\left(Z|X^{(j+1)}, Y^{(j+1)}\right) \end{aligned}$$

Prenons l'exemple de 3 variables aléatoires.

Algorithme2 *Gibbs-Sampling* : G nombre d'itérations

---

1) Initialisation de $\left(X^{(0)}, Y^{(0)}, Z^{(0)}\right)$
2) Pour $g \in \{1, \ , \ldots, \ G\}$
3) Tirage aléatoire de $X^{(g+1)}$ conditionnellement à $\left(X^{(g)}, Z^{(g)}\right)$
4) Tirage aléatoire de $Y^{(g+1)}$ conditionnellement à $\left(Y^{(g+1)}, Z^{(g)}\right)$
5) Tirage aléatoire de $Z^{(g+1)}$ conditionnellement à $\left(Z^{(g+1)}, X^{(g+1)}\right)$
6) Fin for

Application

La réponse immunitaire d'un enfant (variable d'intérêt) est mesurée en concentration d'anticorps dirigé contre un antigène à l'aide de la technique *ELISA* [2]. Cette technique est connue pour avoir des seuils de détection et de saturation ; ce qui fait qu'une partie du signal mesuré n'est pas observé. Cette limite de la technique *ELISA* implique des troncatures basses pour les mesures en dessous du seuil de détection et des troncatures hautes au delà du seuil de saturation. D'autre part les données immunologiques sont caractérisées par beaucoup de variabilité et

[2]Enzyme-Linked ImmunoSorbent Assay

	Valeurs réelles		*Estimations*					
			DL		HDL		SEM	
troncatures	Paramètre	Valeur	Valeur	Biais	Valeur	Biais	Valeur	Biais
10%	α	**1.20**	1.36	-0.16	1.47	-0.27	**1.18**	-0.02
	σ	**1.26**	1.17	0.09	1.14	0.12	**1.28**	-0.02
	μ	**2.52**	2.24	0.28	2.09	0.43	**2.55**	-0.03
	ν	**0.011**	0.026	-0.015	0.03	-0.019	**0.015**	-0.004
20%	α	**1.20**	1.54	-0.34	1.66	-0.46	**-1.16**	-0.04
	σ	**1.26**	1.08	0.18	1.05	0.21	**1.34**	0.1
	μ	**2.52**	2.02	0.5	1.88	0.64	**2.63**	-0.1
	ν	**0.011**	0.027	-0.016	0.03	0.019	**0.015**	-0.004
30%	α	**1.20**	1.78	-0.58	1.82	-0.62	**1.37**	-0.17
	σ	**1.26**	1.00	0.26	0.99	0.27	**1.35**	0.09
	μ	**2.52**	1.77	0.75	1.74	0.78	**2.75**	-0.23
	ν	**0.011**	0.032	-0.021	0.032	-0.021	**0.02**	0.009

TAB. 1 – Comparaison quantitative des valeurs réelles des paramètres avec leur estimation en remplaçant les troncatures (10%, 20%, 30%) par le seuil détection (*DL*) ou la moitié du seuil de détection (*HDL*)

d'hétérogénéité. Enfin l'étude concerne sept réponses anticorps ce qui implique une variable de réponse multidimensionnelle.

L'étude de ces données nécessite la création ou l'extension de méthodes adaptées aux caractéristiques statistiques des données. Il s'agit d'un travail en plusieurs étapes dont la première consiste en la gestion des troncatures de manière à optimiser l'estimation des paramètres d'un modèle à données répétées.

Afin de nous assurer de l'efficacité de l'algorithme SEM pour la gestion des troncatures nous comparons sur des simulations, l'estimation des paramètres par SEM et l'estimation des paramètres avec les méthodes usuelles. Le modèle est un modèle linéaire à effets mixtes avec ordonnée à l'origine et pente aléatoire (2).

Le jeu de données est constitué de 184 enfants de six mesures chacun. $c_{i,t}$ représente la concentration d'anticorps de l'enfant i au temps t. Les mesures correspondent aux temps en mois : $t \in \{3, 6, 9, 12, 15, 18\}$

$$\log c_{i,t} = (\alpha + \alpha_i) + \beta_i t + \varepsilon_{i,t} \tag{2}$$

– $(\alpha_i, \beta_i) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ effets aléatoires ;
– $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$ résidus iid

Sur ce jeu de données ne présentant pas de troncatures nous créons 10%, 20%, 30% de troncatures, ensuite nous comparons les effets fixes, effets aléatoires et résidus des données complètes (sans troncatures) à leurs estimations. Nous considérons dans un premier temps α_i, β_i indépendants et nous nous intéressons à la gestion des troncatures basses. Ce raisonnement reste valable pour les troncatures hautes. Les paramètres du modèle à estimer sont :

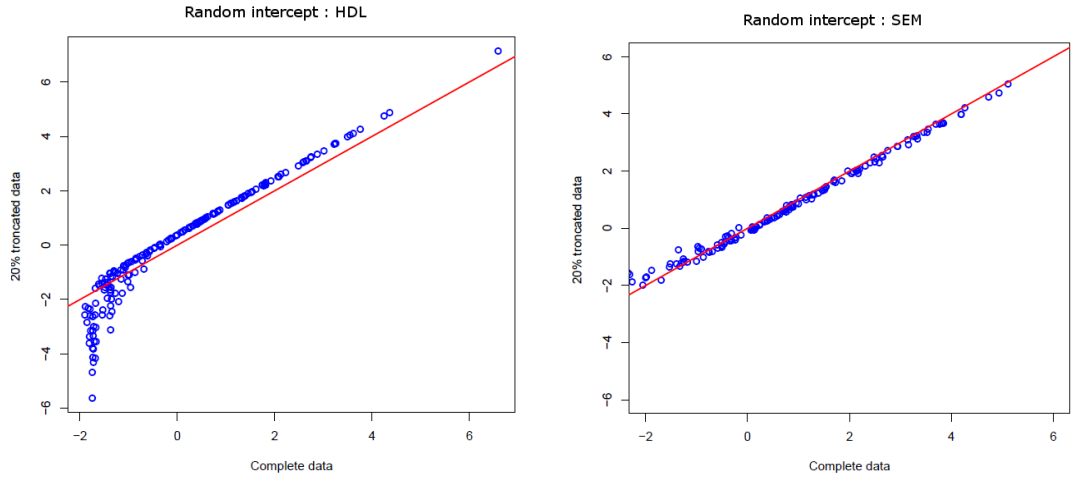


FIG. 1 – Comparaison graphique des valeurs réelles des effets aléatoires pour chaque enfant (en abscisse) avec leur estimation (en ordonnée) en remplaçant les 20% de troncature par la moitié du seuil de détection (*HDL*). La droite rouge est la première bissectrice ; plus on est proche de cette droite et moins il y a de biais au niveau des estimateurs.

- effet fixe : α ;
- variances des effets aléatoires : μ^2, ν^2 ;
- variance résiduelle : σ^2.

L'estimateur converge après (environ) 500 itérations de l'algorithme SEM, avec une phase de *burn-in* de 100 itérations.
En observant le tableau 1, nous remarquons que l'imputation déterministe des troncatures conduit à un biais important, il en est de même pour la suppression des troncatures (pas représenté dans ce tableau). Parmi les quatre méthodes d'estimations testées, l'algorithme SEM se montre plus efficace que les autres quel que soit le seuil de troncature.

La figure 1 représente les estimations des α_i pour chaque enfant. On remarque qu'en remplaçant les troncatures par la moitié du seuil de détection, il apparaît un biais systématique au niveau des estimations qui est incontrôlable. On retrouve ce même biais pour la suppression des troncatures ou le remplacement par le seuil de détection. Les estimations par l'algorithme SEM réduisent amplement ce biais.

Discussion

Dans ce travail nous présentons une méthode qui permet de prendre en compte les données tronquées dans un modèle linéaire à effet mixte. Nous validons cette approche sur des données de réponse anticorps issues de l'étude de cohorte menée au Bénin par l'équipe de l'IRD UMR216, où la concentration d'anticorps représente notre variable d'intérêt.

Nous comparons cette méthode aux autres approches d'imputation simple couramment uti-

lisées en faisant varier le taux de troncature. En plus du biais sur les effets fixes d'un modèle linéaire à effets mixtes, on constate un biais important au niveau de l'estimation des effets aléatoires, or sur l'étude de la réponse anticorps du nouveau-né face au paludisme, ce sont ces effets aléatoires qui permettent de comparer les enfants entre eux et de mieux observer la variabilité liée à chaque individu. L'approche SEM couplée à du *Gibbs-Sampling* réduit nettement ce biais.

Un des inconvénients de cette méthode est le temps de calcul relativement long dû à l'estimation des paramètres d'un modèle linéaire à effets mixtes à chaque itération de l'algorithme. D'autre part se pose la question d'initialisation du paramètre pour converger plus rapidement. Les méthodes directes implémentées par Hughes (1999) et Thiebaut and Jaqmin-Gadda (2004) seront plus optimales en terme de vitesse de convergence dans la mesure où l'estimation est intégrée directement dans le modèle. Notre méthode a toutefois l'avantage d'être facile à implémenter sur des logiciels statistiques standards (R par exemple), de plus elle s'adapte aisément au modèle utilisé (modèle linéaire à effets mixtes, modèles mixtes hiérarchiques). Ce travail constitue une première étape pour la gestion des troncatures de manière simple et plus rigoureuse que les méthodes d'imputation simples usuelles dans le cas unidimensionnel.

Les prochaines étapes de ce travail seraient de tester l'efficacité de cette méthode dans le cadre d'une analyse jointe, puisque dans l'application sur les données immunologiques il y a un aspect multidimensionnel à prendre en compte. D'autre part pour tenir compte de l'hétérogénéité des données, nous passerons à un modèle à trajectoires latentes ou modèle de mélange avec gestion des troncatures.

Références

George Casella and Edward I. George. Explainning the gibbs sampler. *American Statistical association*, 46 :167–174, 1992.

Gilles Celeux and Jean Diebolt. A random imputation principle : the stochastic em algorithm. *Rapport de recherche INRIA*, 901, 1988.

James P. Hughes. Mixed models with censored data with application in hiv infection. *Biometrics*, 55 :625–629, 1999.

Helene Jaqmin-Gadda and al. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 2000.

Adeline Samson and al. Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model : application to hiv dynamics model. *Journal Computational Statistics ans Data Analysis*, 51, 2006.

Rodolphe Thiebaut and Helene Jaqmin-Gadda. Mixed models for longitudinal left-censored repeated measure. *Computer Methods and Programs in Biomedicine*, 74 :255–260, 2004.

JEUDI 26 MAI 2011, 14h

La Statistique Spatiale

Analyse spatiale du travail des enfants, *Sébastien Djienouassi*

Cette étude examine la relation entre offre d'éducation et travail des enfants. Plus particulièrement, nous nous intéressons aux causes de l'échec de la scolarisation universelle face au travail des enfants dans les pays en développement. En utilisant les données de l'enquête sur l'emploi et le secteur informel réalisée par l'Institut National de la Statistique, nous arrivons à la conclusion que cet échec est imputable à l'absence de coordination entre les politiques d'éducation, de santé et de répartition de revenus. De plus, les politiques de lutte contre le travail des enfants ne tiennent pas compte de l'hétérogénéité des comportements et des effets d'autocorrélation spatiale caractérisant les localités et retraçant l'effet de voisinage.

Modelling spatial autocorrelation in hyperspectral remote sensing data, *Saoussen Bahria and Mohamed Limam*

Modeling spatial correlation is a key challenge in classification problem that arise in remote sensing field. Spatial autoregressive model (SAR) is an extension of the classical regression model used for modeling spatial autocorrelation in spatial data. Geostatistical model produce great results for modeling spatial autocorrelation in remote sensing literature. The purpose of this paper is the comparison of the two family of models applied on a benchmark hyperspectral data set. The AVIRIS-Indian Pines 1992 data was used for empirical work. Numerical results show the superiority of geostatistical model to estimate spatial autocorrelation in hyperspectral remote sensing data.

Simulation d'un vecteur gaussien : une approche propagative de l'échantillonneur de Gibbs, *Nicolas Desassis and Christian Lantuéjoul*

Partant de l'échantillonneur de Gibbs, un algorithme itératif est construit, permettant de simuler un vecteur gaussien sans recourir à l'inversion, ni même à la factorisation d'une matrice de covariance. Un aperçu est donné des multiples façons dont cet algorithme peut être implémenté. Son efficacité pratique est illustrée sur un exemple, et un résultat théorique de convergence est énoncé.

Estimation for random coefficient autoregressive models on a plane,

Soumia Kharfouchi and Houda Mehri

In this paper, we introduce a new class of non linear spatial models. This class is generated by a spatial random coefficient autoregressive process. A sufficient condition for the existence of a stationary and ergodic solution is given and a quasi-maximum likelihood estimation procedure is proposed.

A factor model approach for the segmentation of correlated time series,

Emilie Lebarbier and Stéphane Robin

On s'intéresse à la segmentation d'un ensemble de séries d'observations corrélées, typiquement organisées spatialement. On propose de modéliser la dépendance entre les séries au moyen d'un modèle à facteur. Cette modélisation de la dépendance autorise l'utilisation d'algorithmes de segmentation efficaces pour obtenir les estimateur du maximum de vraisemblance. On propose également une procédure de sélection de modèle pour déterminer le nombre de points de ruptures ainsi que le nombre de facteurs.

ANALYSE SPATIALE DU TRAVAIL DES ENFANTS

Sébastien DJIENOUASSI
Ingénieur statisticien
MSs Economie du développement

Institut National de la Statistique
Département des synthèses et analyses économiques
B.P. : 134 INS Yaoundé (Cameroun)
Tél. +237 70839593
djienouassi2006@yahoo.fr / djienouassi@yahoo.fr

Résumé

Cette étude examine la relation entre offre d'éducation et travail des enfants. Plus particulièrement, nous nous intéressons aux causes de l'échec de la scolarisation universelle face au travail des enfants dans les pays en développement. En utilisant les données de l'enquête camerounaise sur l'emploi et le secteur informel réalisée par l'Institut National de la Statistique, nous arrivons à la conclusion que cet échec est imputable à l'absence de coordination entre les politiques d'éducation, de santé et de répartition de revenus. De plus, en utilisant les modèles d'économétrie spatiale, nous arrivons à la conclusion que les politiques de lutte contre le travail des enfants ne tiennent pas compte de l'hétérogénéité des comportements et des effets d'autocorrélation spatiale caractérisant les localités et retraçant l'effet de voisinage.

Mots clés : Travail des enfants ; Education ; Santé ; Inégalités ; Effets de diffusion ; Effets multiplicateurs.

Abstract

This study examines the relationship between education and child labour. In particular, we investigate reasons why the universal schooling has failed against child labor in developing countries. Using data from the Cameroonian survey on employment and informal sector conducted by the National Institute of Statistics, we find evidence that the failure of universal schooling should be imputable to the lack of coordination between education, health and income distribution policies. Furthermore, using models for spatial econometrics, we concluded that policies against child labor do not take into account the heterogeneity in households' behavior and spatial autocorrelation effects which characterize localities and retrace the effect of neighborhood.

Keywords: Child labor; Education; Health; Inequality; Diffusion' effects; Multiplier effects

1. Introduction

Le phénomène du travail des enfants dans le monde reste préoccupant. Selon la dernière estimation mondiale de l'incidence du travail des enfants effectuée par l'Organisation Internationale du Travail (OIT, 2005), plus de 246 millions d'enfants sont au travail dans le monde. L'Afrique subsaharienne présente la proportion la plus élevée avec près d'un tiers des enfants de moins de 14 ans exerçant une activité économique.

La plupart des pays ont adopté des lois interdisant le travail des enfants ou visant à y imposer des restrictions rigoureuses, dont la majeure partie est inspirée et guidée par des normes internationales adoptées par l'OIT. Malgré ces efforts, le travail des enfants continue d'être largement répandu. L'efficacité des politiques économiques visant à réduire le phénomène requiert la connaissance des facteurs qui expliquent son existence. C'est ainsi qu'il est impératif qu'une politique dans ce domaine soit basée sur une analyse rigoureuse du travail des enfants.

Des débats autour des déterminants du travail des enfants abondent de plus en plus dans la littérature théorique et empirique. Au cœur de ces débats se situe en bonne place l'offre d'éducation qui est, d'après le

BIT, l'alternative inéluctable au travail des enfants (BIT, 2004). Bien qu'il soit connaissance commune que l'investissement en capital humain des enfants pour les ménages et l'acquisition de technologies intensives en main d'œuvre qualifiée pour les entreprises soient Pareto-efficients, les parents continuent de mettre les enfants sur le marché du travail. Ceci en raison de la faible offre d'emploi qualifié dans les entreprises et de la réticence des entrepreneurs à investir dans des technologies qualifiées, étant donné qu'ils n'espèrent pas avoir une main d'œuvre qualifiée dans le futur (Dessy et Pallage, 2001). L'intensification des programmes d'éducation serait donc un signal en direction des entreprises sur la disponibilité future du travail qualifié. Dans leur modèle, Dessy et Pallage montrent qu'une scolarisation obligatoire des enfants, combinée à des incitatifs adéquats suffit à éradiquer le travail des enfants. Cette thèse est largement soutenue par les Objectifs du Millénaire pour le Développement (OMD[1]) qui définissent d'ailleurs l'éducation pour tous comme un des objectifs à atteindre d'ici à 2015.

De plus, des études empiriques dont celle de Psacharopoulos (1994) ont montré que les rendements de tous les niveaux d'éducation (publique et privée) sont plus élevés que les rendements des investissements en capital dans toutes les régions du monde. De tout point de vue, l'investissement en éducation est une opportunité attractive. La question qu'on serait donc en droit de se poser est celle de savoir pourquoi le travail des enfants continue t-il de galoper dans les pays en développement alors qu'il est économiquement inefficient. L'éducation universelle aurait-elle échoué face au phénomène de travail des enfants dans ces pays ?

Un pan de la littérature soutient en effet que les vastes dépenses d'éducation observées dans les pays en développement ne sont pas la meilleure façon d'éradiquer le travail des enfants. L'offre d'éducation serait efficace, soit si elle était couplée à l'offre de santé dans un environnement où les ménages ont une forte aversion au risque (Chakraborty et Das, 2005 ; Grigoriou et Graziosi, 2008), soit dans un environnement où les inégalités de distribution de revenus sont faibles (Tanaka, 2003), soit dans un contexte de marché de crédit parfait (Ranjan, 1999).

Mais toutes ces études manquent de rigueur dans le traitement du travail des enfants en tant que phénomène social et par conséquent, sujet à des effets multiplicateurs des pairs et à des effets de diffusion. En effet, la pauvreté, largement admise comme cause principale du travail des enfants, a été démontrée par de nombreux auteurs (Elbers, C et al., 2005 ; Benson et al., 2005) comme sujette à une forte concentration spatiale positive. Il y a donc de fortes raisons de penser que le travail des enfants le soit aussi.

Notre étude se situe à la croisée de ce débat. Le but est de comprendre sur une base empirique pourquoi l'éducation universelle aurait échoué face au travail des enfants. Nous analyserons en particulier l'effet sur le travail des enfants, de l'interaction entre offre d'éducation, offre de santé et répartition de revenus. La cohabitation entre gratuité de l'enseignement primaire, fortes inégalités de répartitions de revenus et forte incidence du travail des enfants suscite la curiosité et fait de l'expérience camerounaise un cadre judicieux d'analyse pour comprendre et tester le sens des corrélations ci-dessus. La particularité de notre étude est la prise en compte de l'interaction spatiale du travail des enfants. Les données utilisées sont issues de l'enquête nationale réalisée sur l'emploi et le secteur informel en 2005 par l'Institut National de la Statistique.

2. Approche méthodologique

La majorité des études sur le travail des enfants au Cameroun arrivent à la conclusion que, comme la pauvreté, le travail des enfants est prépondérant dans les régions septentrionales du pays. Une confrontation avec l'offre d'éducation montre que dans les départements où le nombre d'instituteurs pour 100 élèves est très faible, le travail des enfants est en général très élevé. Le besoin d'intégrer la dimension spatiale pour mesurer l'effet de l'offre d'éducation comme politique de ciblage contre le travail des enfants s'impose. Cependant, la prise en compte de cette dimension ne peut pas se limiter à l'introduction de quelques variables supplémentaires, il faut tenir compte de l'effet de voisinage et de l'hétérogénéité des comportements des ménages des différentes localités. Le recours aux méthodes de l'économétrie spatiale est une nécessité. Les estimations intégrant ces nouvelles dimensions donneraient lieu à des meilleures approximations de la réalité et une meilleure évaluation des effets des différentes politiques de ciblage. Nous allons de ce fait tester le problème d'autocorrélation (interaction liée à la proximité) et d'hétérogénéité spatiale. Le test

[1] Les Objectifs du millénaire pour le développement (OMD) sont huit objectifs que les États membres de l'ONU ont convenus d'atteindre d'ici à 2015. La déclaration fut signée en septembre 2000.

d'hétérogénéité se fera par la prise en compte des variables telles que le nombre d'unités de consommation[2], la structure de l'économie (proportion d'agriculteurs), etc.

2.1 Le test d'autocorrélation spatiale

La première étape de notre analyse exploratoire consiste à évaluer l'autocorrélation spatiale globale au sein de l'échantillon, afin de déterminer si, globalement, il existe une concentration spatiale des localités similaires en termes de travail des enfants et des autres variables explicatives. Afin de tester cette hypothèse, nous faisons recours à deux statistiques : la statistique I de Moran et la statistique c de Geary définies respectivement comme suit (Anselin 1995) :

$$I = \frac{N\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{P\sum_i (y_i - \bar{y})^2}$$

$$c = \frac{(N-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2P\sum_i (y_i - \bar{y})^2}$$

Où y_i est la valeur de la variable observée dans la localité i, $\bar{y}$ est la moyenne de cette variable au niveau national, N est le nombre de localités, $P = \sum_{i \neq j} w_{ij}$ est la somme des pondérations et w_{ij} est l'élément de la matrice de poids spatial[3]. Une valeur positive et statistiquement significative de ces statistiques indique la présence d'une concentration spatiale des valeurs similaires, alors qu'une valeur négative et statistiquement significative indique une concentration de valeurs dissimilaires.

Tableau .1 : Test d'autocorrélation spatiale

	Moran		Geary	
	I	p-value	c	p-value
Travail	0.119*	0.061	0.809**	0.034
Gini	0.218***	0.004	0.717***	0.008
Education	0.068	0.134	0.715*	0.079
Santé	0.025	0.304	0.864	0.144
Pauvreté	0.524***	0.000	0.472***	0.000

Significativité : 10% (*), 5% (**), 1% (***)
Source : nos soins à partir des données d'EESI

Le tableau 1 donne les résultats des deux tests de Moran et de Geary pour les variables taux de travail des enfants (travail), indice de Gini pour la mesure des inégalités (Gini), nombre d'instituteurs pour 100 élèves (Education), nombre de cadres moyens et supérieurs de santé pour 1000 habitants (Santé) et le taux de pauvreté par région (Pauvreté). Mises à part les variables Education et Santé qui dépendent beaucoup plus des politiques publiques, les deux tests sont positivement significatifs pour Travail, Gini et Pauvreté. Ce qui indique la présence d'une autocorrélation spatiale (relation avec le décalage spatial) pour ces variables. Nous pouvons donc confirmer l'existence des concentrations de pauvres, à revenus très inégalement répartis et à forte incidence du travail des enfants, et des concentrations de non pauvres, à faibles inégalités de répartition de revenus et faible incidence du travail des enfants. Ce qui retrace la significativité de l'effet de voisinage et de concentration dont la non prise en compte pourrait justifier l'échec des programmes visant à éradiquer le travail des enfants. Ce constat nous pousse à tenir compte, dans nos procédures d'estimations économétriques, de l'autocorrélation et de l'hétérogénéité spatiale.

2.2 Méthodes d'estimation

Comme dans l'article de Elbers et al (2005), nous utilisons une équation de la forme :

$$Y_i = X_i\beta + \varepsilon_i$$

[2] Les unités de consommations sont standardisées grâce à l'échelle RDA. Plus loin les coefficients de pondération sont présentés.

[3] La matrice de poids spatial est une matrice carrée d'ordre N, symétrique et positive. L'élément w_{ij} mesure le degré d'interaction entre deux régions voisines *i* et *j. Celle que nous utilisons ici et qui est la* plus utilisée est la matrice de contiguïté d'ordre 1, son élément w_{ij} est égal à 1 si les deux régions *i* et *j* ont une frontière commune, il est égal à zéro dans le cas contraire

Où Y_i est le taux de travail des enfants de la région i, X est la matrice des variables explicatives et ε_i est le terme d'erreur qui suit la loi normale. Deux méthodes d'estimation seront mises à contribution afin d'estimer les paramètres de cette équation.

Modèle linéaire

Nous estimons dans un premier temps le taux d'activité des enfants des régions par les moindres carrés ordinaires. Dans cette estimation nous faisons abstraction de l'autocorrélation spatiale qui peut caractériser le comportement des ménages. En outre nous supposons la stationnarité des effets des différentes variables explicatives sur l'offre microéconomique du travail des enfants.

Modèle spatial

L'introduction de l'interaction spatiale dans le modèle ci-dessus peut prendre deux formes. Dans la première, on considère que l'interaction spatiale porte sur la variable expliquée. On aboutit alors au modèle spatial autorégressif (SAR) :

$$y_i = \rho W y_i + X_i\beta + \varepsilon_i \Leftrightarrow (I - \rho W)y_i = X_i\beta + \varepsilon_i$$

où, on fait sur ε les hypothèses standard des moindres carrés, E(ε) = 0 et V(ε) = σ^2V. Cette méthode insère une variable spatiale retardée parmi les variables explicatives définie par Wy où W est la matrice des effets d'autorégression spatiale (ici nous utilisons la matrice spatiale de contiguïté d'ordre 1). Le modèle spatial autorégressif s'impose en particulier dès qu'on n'a aucune raison de penser que la variable expliquée est d'espérance nulle partout dans l'espace (Jayet, 2001).

Dans le deuxième cas, l'interaction spatiale porte sur la partie aléatoire du modèle, ε, qui suit un processus autorégressif spatial. On aboutit au modèle avec autocorrélation spatiale des résidus (SEM) :

$$\left.\begin{array}{l} y_i = X_i\beta + \varepsilon_i \\ (I - \gamma G)\varepsilon_i = \eta_i \end{array}\right| \Leftrightarrow (I - \gamma G)(y_i - X_i\beta) = \eta_i$$

où G est la matrice des effets d'autocorrélation spatiale et η est un vecteur d'aléas indépendants vérifiant E(η) = 0 et V(η) = σ^2V.

Grâce à ces estimations nous pourrons tenir compte de l'effet de voisinage et d'interdépendance entre les différentes régions mesurées par le coefficient autorégressif ρ ou le coefficient d'autocorrélation γ.

3. Principaux résultats

Le tableau 2 nous mène aux conclusions similaires à celles des équations de participation et d'offre du travail des enfants que nous ne présentons pas ici. Il ressort que l'augmentation du nombre d'instituteurs (INSTITUT), l'accroissement du nombre moyen d'unités de consommation par ménage (UCMOY), l'augmentation des salaires des instituteurs conjointement à l'amélioration de l'offre de santé (LNSAL*SANTE), l'augmentation des salaires des instituteurs dans un environnement de fortes inégalités (LNSAL*GINI) ont tous des effets positifs et significatifs sur l'éradication du travail des enfants. Par contre l'augmentation du salaire des instituteurs (LNSAL), l'accroissement des inégalités (GINI), la proportion des adultes analphabètes (ANALPH), l'accroissement de l'offre de santé (SANTE) ont tendance à perpétuer le phénomène. Il est important de noter que l'offre d'éducation à travers les constructions massives des écoles n'a pas d'effet significatif sur le travail des enfants. Le coefficient devant la variable distance parcourue pour atteindre l'école publique la plus proche (DISECOLE) est non significatif dans toutes les estimations, confortant ainsi l'idée selon laquelle les vastes dépenses d'éducation observées récemment dans les pays en développement, ne sont pas un bon moyen pour l'éradication du travail des enfants.

De façon générale, les signes associés aux différentes variables explicatives sont, pour la plupart, conformes à l'intuition, cependant, les grandeurs rapportées par les différentes méthodes d'estimation sont assez différentes. Une première comparaison des coefficients estimés des modèles MCO (colonnes 1 et 4) et des modèles SAR (colonnes 2 et 5) montre que les variables INSTITUT LNSAL SANTE, INSTITUT*SANTE, INSTITUT*GINI, LNSAL*SANTE, LNSAL*GINI, ANALPH, UCMOY sont toutes significatives dans les deux types d'estimation. Cependant, les variables des inégalités (GINI) et de temps mis pour atteindre l'école publique la plus proche (DURECOLE) ne sont significatives que dans les modèles SAR. De plus, l'effet des

variables dans les modèles MCO a en général une ampleur plus importante. Cette surévaluation pourrait être la conséquence du cumul de l'effet direct et l'effet indirect de chaque variable explicative.

La significativité du coefficient de corrélation rho justifie l'utilisation du modèle spatial autorégressif. La significativité de rho signifie qu'en moyenne la proportion d'enfants actifs dans une localité n'est pas seulement expliquée par les valeurs des variables explicatives associées à cette localité, mais aussi par celles associées à toutes les localités voisines à travers la transformation spatiale inverse $(I-\rho W)^{-1}$. Cet effet de multiplicateur spatial décline avec l'éloignement.

Tableau 2 : Résultats des estimations spatiales

TRAVAIL	10 – 17 ans			10 – 14 ans		
	(1) MCO	(2) SAR	(3) SEM	(4) MCO	(5) SAR	(6) SEM
INSTITUT	-0,295***	-0,284***	-0,296***	-0,247***	-0,234***	-0,267***
	(0,048)	(0,040)	(0,044)	(0,065)	(0,056)	(0,064)
LNSAL	0,252***	0,230***	0,252***	0,257***	0,232***	0,261***
	(0,068)	(0,048)	(0,048)	(0,084)	(0,061)	(0,063)
GINI**	1,511	1,489*	1,433**	1,603	1,515	1,218**
	(1,065)	(0,810)	(0,572)	(1,229)	(0,940)	(0,618)
SANTE	0,341***	0,374***	0,370***	0,253**	0,275***	0,277**
	(0,107)	(0,091)	(0,105)	(0,111)	(0,094)	(0,123)
INSTITUT*SANTE	0,013***	0,014***	0,013***	0,014***	0,014***	0,012***
	(0,005)	(0,004)	(0,004)	(0,005)	(0,004)	(0,004)
INSTITUT*GINI	0,634***	0,613***	0,642***	0,505***	0,483***	0,552***
	(0,105)	(0,089)	(0,103)	(0,150)	(0,133)	(0,147)
LNSAL*SANTE	-0,039***	-0,042***	-0,042***	-0,031***	-0,033***	-0,033***
	(0,010)	(0,009)	(0,010)	(0,011)	(0,009)	(0,012)
LNSAL*GINI	-0,431***	-0,409***	-0,428***	-0,383***	-0,358***	-0,398***
	(0,067)	(0,056)	(0,058)	(0,094)	(0,083)	(0,084)
ANALPH	0,533***	0,595***	0,478***	0,486***	0,549***	0,456***
	(0,137)	(0,102)	(0,083)	(0,171)	(0,123)	(0,100)
UCMOY	-0,127***	-0,113***	-0,125***	-0,127***	-0,108***	-0,132***
	(0,037)	(0,031)	(0,025)	(0,043)	(0,035)	(0,032)
DISECOLE	-0,006	-0,008	0,026	-0,008	-0,008	0,013
	(0,059)	(0,037)	(0,047)	(0,073)	(0,047)	(0,051)
DURECOLE	0,005	0,005*	0,001	0,007	0,006*	0,002
	(0,005)	(0,003)	(0,004)	(0,006)	(0,004)	(0,005)
_cons	-0,415	-0,259	-0,359	-0,818	-0,623	-0,479
	(1,215)	(0,870)	(0,686)	(1,448)	(1,030)	(0,728)
/rho		-0,269**			-0,295*	
		(0,127)			(0,158)	
/lambda			-0,799**			-0,862*
			(0,405)			(0,452)
/sigma		0,057***	0,053***		0,075***	0,068***
		(0,006)	(0,006)		(0,009)	(0,009)
R2	0,88	0,90	0,87	0,81	0,83	O,80

Note : TRAVAIL Taux d'activité des enfants par strate ; INSTITUT Nombre d'instituteurs pour 100 élèves du primaire ; LNSAL Log du salaire moyen des instituteurs par strate ; PLURI Taux de pluriactivité des instituteurs ; GINI** Indice de Gini instrumenté ; SANTE Nombre de cadres de santé (médecins, infirmiers et pharmaciens) pour 1000 habitants ; ANALPH Proportion des adultes sans instruction ; UCMOY Nombre moyen d'unités de consommation par strate ; DISECOLE Distance moyenne par rapport à l'école publique la plus proche ; DURECOLE Temps moyen pour arriver à l'école publique la plus proche par le moyen de locomotion usuel. Entre parenthèses les écarts-types robustes à l'hétéroscédasticité. Significativité : *** p<0.01, ** p<0.05, * p<0.1
Source : nos soins à partir des données d'EESI

La comparaison des coefficients donnés par MCO et par SEM (Spatial Error Model) montre également quelques différences avec notamment, le coefficient des inégalités qui n'est significatif que dans les modèles SEM. La significativité de lamda qui est le coefficient d'autocorrélation spatiale des erreurs traduit le fait qu'un choc aléatoire dans une localité donnée affecte non seulement le travail des enfants de cette localité, mais a également un impact sur le travail des enfants dans les autres localités à travers la transformation $(I-\lambda W)^{-1}$. C'est l'effet de diffusion, effet qui décline aussi avec l'éloignement.

Ces constatations nous autorisent à conclure que la proposition de n'importe quelle politique d'éradication du travail des enfants, et plus particulièrement les politiques de ciblage d'éducation doit prendre en compte l'autocorrélation spatiale du phénomène de travail des enfants. Par conséquent, l'effet de cette politique, comme de n'importe quelle autre politique visant à éradiquer le travail des enfants, ne peut se faire sur la

base d'une estimation basée sur les méthodes a-spatiales classiques qui supposent l'absence d'effet indirect lié au positionnement géographique. L'utilisation des méthodes a-spatiales est d'autant plus contraignante qu'elles ne tiennent pas compte des différences dans les comportements et par suite des réactions des ménages appartenant à des localisations géographiques différenciées. La non prise en compte de ces aspects peuvent en partie, justifier l'échec de l'éducation universelle face au travail des enfants.

4. Conclusion

Les questions portant sur le travail des enfants restent une préoccupation majeure dans les pays en développement au regard des efforts fournis de part et d'autre pour réduire ce phénomène à défaut de l'éradiquer totalement.

Il ressort de ce qui précède que la diversité spatiale est un phénomène important dans l'analyse du travail des enfants. Nos analyses ont été consolidées par la prise en compte des effets de voisinages spatiaux et de l'hétérogénéité des comportements entre les différentes localités géographiques. Nous avons montré que le travail des enfants est un phénomène autorégressif et à erreurs autocorrélées. C'est-à-dire que la proportion d'enfants actifs dans une localité n'est pas seulement expliquée par les valeurs des variables explicatives associées à cette localité, mais aussi par celles associées à toutes les localités voisines et que tout choc aléatoire dans une localité donnée affecte non seulement le travail des enfants de cette localité, mais également le travail des enfants dans les localités voisines. Ces effets diminuent cependant avec l'éloignement. Ce résultat est très fort et explique en partie l'échec de la scolarisation universelle et des vastes programmes d'éducation face au travail des enfants. Toute politique en termes de lutte contre le travail des enfants mérite une attention particulière ; le planificateur devrait tenir compte de l'hétérogénéité des comportements mais aussi des effets d'autocorrélation spatiale caractérisant les localités et retraçant l'effet de voisinage.

5. Quelques Références bibliographiques

[1] Anselin, L., (1995). "Local Indicator of Spatial Association LISA," Geographical Analysis, 27, 93-115.

[2] Basu K et Van P.H. (1998), "The Economics of Child labour", *American Economic Review*, Vol. 88, N°3, pp 412-427

[3] Benson, T., Chamberlin, J., and Rhinehart, I., (2005). "An investigation of the spatial determinants of the local prevelance of poverty in rural Malawi". Food Policy, 30, 532-550.

[4] BIT (2004), « Le travail des enfants : un manuel à l'usage des étudiants », Genève.

[5] Cadiou A. (2002), « Le travail des enfants », Mémoire de DEA de Droit Privé, Mention Sciences Judiciaires et Criminelles, Université de Nantes, 92 pages.

[6] Chakraborty S. and Das M., (2005). "Mortality, fertility, and child labor," Economics Letters 86, 273–278.

[7] Dessy S.E.and Pallage S., (2001). "Child labor and coordination failures," Journal of Development Economics 65, 469–476.

[8] Elbers, C., jean, L., and Peter, L., (2005). "Imputed welfare estimates in regression analysis," Journal of Economic Geography, 5, 101-118.

[9] Grigoriou C. and Graziosi R. G. (2008) "Working versus schooling: the impact of social expenditure", Louvain Economic Review, vol. 74, n° 1, pp. 33-52.

[10] Jayet H. (2001). "Econométrie et données spatiales : Une introduction à la pratique, " Cahiers d'économie et sociologie rurales, n° 58-59

[11] OIT (2005), « Plaidoyer pour des emplois décents aux jeunes », 93ème Conférence de l'Organisation Internationale du Travail, Génève.

[12] Psacharopoulos, G., (1994). "Returns to investment in education: a global update," World Development 22 (9), 1325–1343.

[13] Ranjan P. (1999), "An Economic Analysis of Child Labour", Economics Letters, 1999, Vol 69, pp 99-105.

[14] Tanaka R., (2003). "Inequality as a determinant of child labor," Economics Letters 80, 93–97.

Modeling spatial autocorrelation in hyper-spectral remote sensing data

Saoussen Bahria and Mohamed Limam

LARODEC, High Institute of Management, University of Tunis, Tunisia

Abstract.

Modeling spatial correlation is a key challenge in classification problem that arise in remote sensing field. Spatial autoregressive model (SAR) is an extension of the classical regression model used for modeling spatial autocorrelation in spatial data. The purpose of this paper is the comparison of the SAR model and the linear regression model which ignore the spatial information. The AVIRIS-Indian Pines 1992 data was used for empirical work. Numerical results show the superiority of SAR model to estimate spatial autocorrelation in hyperspectral remote sensing data.

Keywords. Hyper-spectral data, spatial autocorrelation, spatial autoregressive model.

Résumé.

La modélisation de la corrélation spatiale présente un grand défi pour l'amélioration de la classification des données de télédétection. Le modèle autorégressif spatial est une extension du modèle linéaire classique. Ce modèle est utilisé pour la modélisation de l'auto-corrélation spatiale dans les données spatiales. L'objectif de cette communication est la comparaison du modèle autorégressif spatial et le modèle de régression linéaire qui ignore la corrélation spatiale caractérisant les données de télédétection. La performance des deux modèles a été évaluée sur la base de données hyper-spectrale benchmark AVIRIS-Indian Pines 1992 . Les résultats numériques montrent la supériorité du modèle autorégressif spatial pour l'estimation de la corrélation spatiale dans les données hyper-spectrales.

1- Introduction

In spatial literature, classical data mining algorithms, such as linear regression suppose that the learning samples are independently and identically distributed (i.i.d). However, spatial data characterized by a spatial dependence does not respect this assumption. In this context, many approaches propose the improvement of modeling spatial data taking into account spatial autocorrelation. In particular, in the remote sensing field, Griffith (2001) gives a new approach to modeling spatial dependence in high spatial resolution hyper-spectral data.. Licshtein et al (2002) deal with spatial autocorrelation and spatial autoregressive models in ecology. Overmars et *al.* (2003) analyse the spatial autocorrelation in multi-scales land use models. Vanoort et *al.* (2004) introduce spatial variability for the improvement of classification accuracy of agricultural crops in the Duch national land cover database. Dormann et *al.* (2007) gives a review of methods that account spatial autocorrelation in the analysis of species distributional data.

The purpose of this work is the improvement of spatial autocorrelation modeling in hyper-spectral data using SAR model. Linear model assume the hypothesis of independently and identically distributed data. However, the main characteristic of hyper-spectral data is the dependence between neighbor pixels. As a consequence, we compare SAR model versus linear regression one for modeling hyper-spectral data.

This paper is organized as follows: section 2 presents the data and the methodology used is this work. Empirical results are discussed in section 3. Section 4 presents the conclusion and discussion of the proposed work.

2. Data and Methodology

2.1. Dataset: The AVIRIS Indian Pines-1992

The AVIRIS Indian Pines hyperspectral benchmark dataset is available online at http://aviris.jpl.nasa.gov/html/aviris.freedata.html. The scene used for experiment is taken on NASA ER2 flight on 12 June 1992. It consists of 145*145 pixels (20 m of spatial resolution) and 220 spectral bands, with about two-thirds agriculture, and one-third forest or other perennial vegetation, 16 ground truth classes (1-Alfalfa, 2- Corn-notill, 3- Corn-min, 4- Corn,

5- Grass/Pasture, 6- Grass/Trees,7- Grass/Pasture-mowed, 8-Hay/windrowed, 9-Oats, 10-Soybeans-notill, 11- Soybeans-min, 12- Soybeans-clean, 13-Wheat, 14-Woods, 15-Bldg-Grass-Trees, 16-Stone-steel-towers). It is called as Indian Pines 1 in the AVIRIS JPL repository. A total of 20 bands [104-108, 150-163, 220] are removed due to the water absorption region. The remaining 200 spectral channels were effectively used for the experimental study. A total of 10148 and 2031 pixels are used for training and testing, respectively.

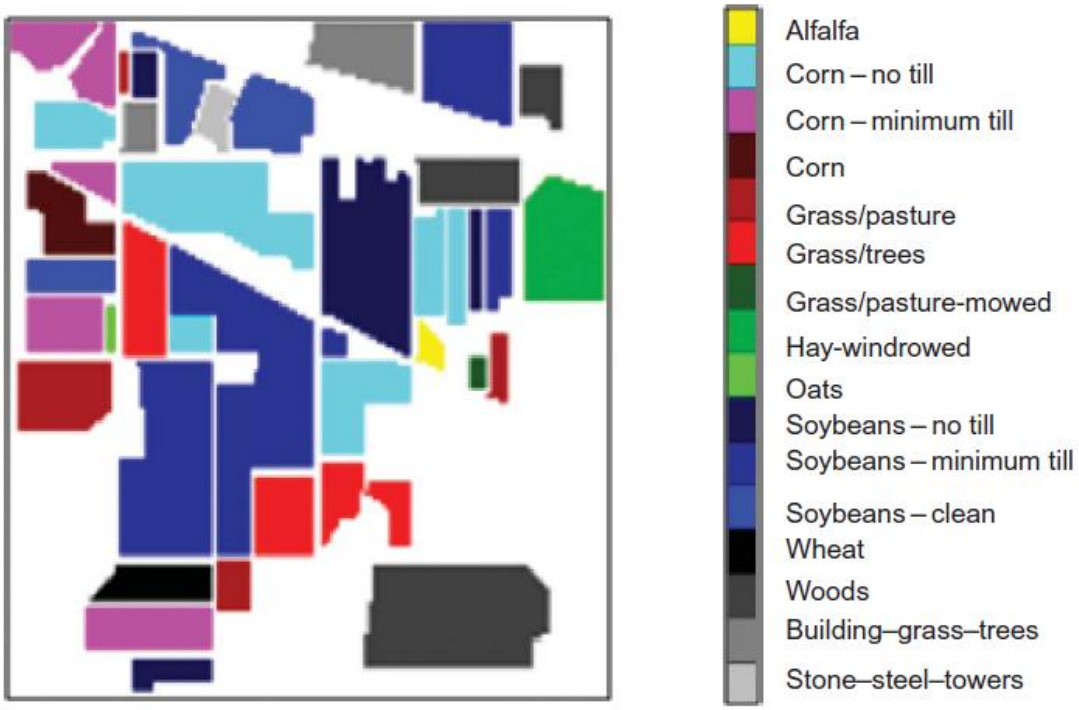


Figure 1. AVIRIS Indian Pines image (Ground reference)

2.2. *Methodology*

A comparative study was conducted. The SAR model was applied on the AVIRIS benchemark data sets. Then linear regression model was used for the same data set. The comparison was done in terms of the pseudo-R^2.

The SAR model is given by:

$$Y = \delta Wy + X\beta + u \qquad (1)$$

$$u = \rho Mu + \varrho$$

where Y is the N × 1 vector of observations on the dependent variable, X is the N × k matrix of observations on the independent variables, W and M are N × N spatial-weighting matrices that parameterize the distance between neighborhoods, u are spatially correlated residuals and ϱ are independent and identically distributed disturbance, $\rho \; and \; \delta$ are scalars that measure, respectively, the dependence of y_i on nearby y and the spatial correlation in the errors.

For the experimental implementation, the software Matlab 6.5 was used on a Pentium M 2.1 GHz personal computer with 3 GB of memory.

3- Results

The superiority of SAR model versus the linear regression model for the estimation of spatial autocorrelation in hyperspectral data was show in table 1. For all selected bands, SAR model gives better Pseudo R^2 than linear regression model. The difference of performance between the two models is very larger, (0.9037-0.5316) for the band 5, (0.8789-0.4722) for the band 60, and (0.8622-0.4510) for the band 120.

Table 1. SAR and linear regression model estimation results for selected AVIRIS Hyper-spectral image's bands

Band	SAR ρ^	Pseudo R^{2-} SAR	Pseudo R –linear model
5	0.9584	**0.9037**	0.5316
60	0.9460	**0.8789**	0.4722
120	0.9390	**0.8622**	0.4510

4- Conclusion

In this work, the spatial autocorrelation is taking into account in hyper-spectral data using SAR model. Empirical results show the higher performance of SAR model vesus linear regression model which ignore the spatial autocorrelation.

Further improvement of hyperspectral data modeling could be investigated based on geostatistical tools…

References.

[1] Dormann, C.F, McPherson, J.K, Arau´ jo. M.B, Bivand, R., Bolliger, J. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30: 609-628, 2007.

[2] Griffith, A.D. (2001). Modeling spatial dependence in high spatial resolution hyperspectral data sets. *Journal of geographical systems*, 4:43–51.

[3] Licshtein, W.J., Simons, R.T, Shriner, S.A., Franzreb, K.E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs,* 72: 445–463.

[4] Overmars, K.P, Koning. G.H.J., and Veldkamp, A. (2003). Spatial autocorrelation in multi-scales land use models. *Ecological Modelling*, 164 : 257–270.

[5] Vanoort, P.A.J, Bregt, A.K, Bruin, S, De Wit, A. J. W. (2004). Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. *International Journal of geographical information science*, 18: 611-626.

Simulation d'un vecteur gaussien: Une approche propagative de l'échantillonneur de Gibbs

Nicolas Desassis & Christian Lantuéjoul

Ecole des Mines-Paristech, 35 rue Saint-Honoré, 77305 Fontainebleau, France

Résumé: Partant de l'échantillonneur de Gibbs, un algorithme itératif est construit, permettant de simuler un vecteur gaussien sans recourir à l'inversion, ni même à la factorisation d'une matrice de covariance. Un aperçu est donné des multiples façons dont cet algorithme peut être implémenté. Son efficacité pratique est illustrée sur un exemple, et un résultat théorique de convergence est énoncé.

Abstract: Starting from the Gibbs sampler, an iterative algorithm is designed for simulating a Gaussian vector, that requires neither the inversion nor the factorization of a covariance matrix. A brief survey is given of the various ways to implement it. An example illustrates its feasibility, and a theoretical result is stated about its rate of convergence.

Mots clés: vecteur aléatoire gaussien, échantillonneur de Gibbs, simulation

De l'échantillonneur de Gibbs à un algorithme de propagation

Soit $Y = (Y_\alpha, \alpha \in I)$ un vecteur gaussien centré normé de matrice de covariance C. Pour simuler Y itérativement, une approche classique est l'échantillonneur de Gibbs des frères Geman (1984). Dans sa formulation géostatistique, une itération à partir de l'état courant y^c se fait comme suit: une composante α est d'abord sélectionnée au hasard, puis la valeur courante y^c_α est remplacée par une nouvelle valeur y^n_α générée selon une loi gaussienne $\mathcal{G}(y_\alpha(I^\alpha), \sigma^2_\alpha)$, les autres valeurs restant inchangées, i.e. $y^n_{-\alpha} = y^c_{-\alpha}$. Dans cette loi gaussienne, $y_\alpha(I^\alpha) = \sum_{\beta \neq \alpha} \lambda^\alpha_\beta y^c_\beta$ désigne le krigeage simple de la composante α sur les autres composantes, et σ^2_α la variance de krigeage. L'ensemble de tous les pondérateurs de krigeage est donné par l'inverse de la matrice de covariance de Y:

$$\lambda^\alpha_\beta = -\frac{C^{-1}_{\alpha\beta}}{C^{-1}_{\alpha\alpha}}$$

Dès que le vecteur Y possède un assez grand nombre de composantes, le calcul de la matrice C^{-1} devient impossible et l'échantillonneur de Gibbs ne peut plus être rigoureusement implémenté.

Une façon de contourner le problème est de travailler non pas sur Y, mais sur $X = C^{-1}Y$. Comme le remarquent Galli et Gao (2001), X est un vecteur gaussien de matrice de

covariance C^{-1} ce qui fait que l'échantillonneur de Gibbs peut être appliqué sur ce vecteur sans aucune approximation. Il suffit ensuite de remonter à Y en posant $Y = CX$.

De façon plus précise, le krigeage de X_α sur les autres variables de X vaut

$$x_\alpha(I^\alpha) = -\sum_{\beta\neq\alpha} \frac{C_{\alpha\beta}}{C_{\alpha\alpha}} x^c_\beta = -\sum_{\beta\neq\alpha} C_{\alpha\beta} x^c_\beta = x^c_\alpha - y^c_\alpha$$

et la variance de krigeage vaut 1. On a donc $x^n_\alpha = x^c_\alpha - y^c_\alpha + u$, où u est une valeur gaussienne centrée réduite, ainsi que $x^n_{-\alpha} = x^c_{-\alpha}$.

Regardons maintenant ce que devient ce passage de x^c à x^n sur le vecteur d'intérêt Y:

$$y^n_\beta = \sum_\gamma C_{\beta\gamma} x^n_\gamma = C_{\beta\alpha}(x^c_\alpha - y^c_\alpha + u) + \sum_{\gamma\neq\alpha} C_{\beta\gamma} x^n_\gamma = y^c_\beta + C_{\beta\alpha}(-y^c_\alpha + u) \qquad \beta \in I$$

Pour $\beta = \alpha$, cela donne $y^n_\alpha = y^c_\alpha - y^c_\alpha + u = u$. Finalement

$$y^n_\beta = y^c_\beta + C_{\alpha\beta}(y^n_\alpha - y^c_\alpha) \qquad \beta \in I \tag{1}$$

L'algorithme obtenu s'interprète ainsi: à chaque itération, un indice α est sélectionné au hasard. La valeur courante y^c_α est remplacée par une valeur gaussienne centrée réduite y^n_α. L'écart entre y^n_α et y^c_α est ensuite *propagé* aux autres composantes. Plus $C_{\alpha\beta}$ est faible, plus l'effet de la propagation sur la composante β est limité. Une écriture pseudocode de cet algorithme est donnée ci-dessous:

(i) poser $y^c = 0$*;*
(ii) générer $\alpha \sim \mathcal{U}(I)$ *ainsi que* $y^n_\alpha \sim \mathcal{G}$*;*
(iii) poser $y^n_\beta = y^c_\beta + C_{\alpha\beta}(y^n_\alpha - y^c_\alpha)$ *pour tout* $\beta \neq \alpha$*;*
(iv) retourner en (ii).

Pour la suite, la composante α sera appelée *pivot*.

Généralisations

Un aspect attrayant de cet algorithme propagatif est qu'il est susceptible de multiples généralisations.

Une première généralisation consiste à générer la variable pivot y^n_α en fonction de la valeur courante y^c_α. L'exemple le plus simple est de prendre $y^n_\alpha = r y^c_\alpha + \sqrt{1-r^2}u$, où $u \sim \mathcal{G}$ et $|r| < 1$. Dans le cas $r = 0$, on retrouve le cas de la section précédente. Pour $r \approx 1$, cette mise à jour de la variable pivot garantit que l'état nouveau reste proche de l'état courant, ce qui est parfois bien utile lorsque l'algorithme propagatif est utilisé pour spécifier un état candidat qui doit être ensuite validé selon un critère de Metropolis-Hastings: un écart trop important entre les états courant et nouveau conduit à un taux de rejet inacceptable. Il en est ainsi de la simulation conditionnelle de certains modèles stochastiques spatiaux (Lantuéjoul, 2011).

Pour la seconde généralisation, écrivons l'équation de propagation (1) sous la forme

$$y^n_\beta - C_{\alpha\beta} y^n_\alpha = y^c_\beta - C_{\alpha\beta} y^c_\alpha, \qquad \beta \in I.$$

Compte tenu de ce que $C_{\alpha\beta} y^c_\alpha$ et $C_{\alpha\beta} y^n_\alpha$ sont les krigeages de y^c_β et de y^n_β sur les valeurs pivot y^c_α et y^n_α, il apparait que l'équation (1) préserve les résidus de krigeage.

La plupart du temps, le krigeage n'est pas effectué à partir d'une seule donnée α, mais à partir d'un ensemble plus riche A. A partir de là, on peut se demander si l'algorithme propagatif reste valable lorsque le pivot α est remplacé par une famille A de pivots, l'étape de propagation étant conçue de façon à préserver les résidus de krigeage. On est donc amené à considérer l'algorithme suivant, dans lequel $y^c_\beta(A)$ et $y^n_\beta(A)$ désignent les krigeages de y^c_β et de y^n_β sur les valeurs des pivots y^c_A et y^n_A, C_{AA} est la matrice de covariance du vecteur aléatoire $Y_A = (Y_\alpha, \alpha \in A)$, et $\mathcal{S}$ est une loi de probabilité qui sélectionne une partie non vide de I:

(i) poser $y^c = 0$*;*
(ii) générer $A \sim \mathcal{S}(I)$ *ainsi que* $y^n_A \sim \mathcal{G}(0, C_{AA})$*;*
(iii) poser $y^n_\beta = y^c_\beta + y^n_\beta(A) - y^c_\beta(A)$ *pour tout* $\beta \notin A$*;*
(iv) retourner en (ii).

La convergence de cet algorithme vers la loi gaussienne cible est établie moyennant une légère restriction sur le choix de la loi $\mathcal{S}$ des pivots: chaque composante α doit avoir une chance non nulle de figurer dans un pivot (Lantuéjoul, 2011).

Bien sûr, il n'y a aucun inconvénient à combiner les deux généralisations précédentes. On peut ainsi générer y^n_A en fonction de y^c_A, par exemple en posant

$$y^n_A = r y^c_A + \sqrt{1 - r^2} u_A$$

où u_A est un vecteur gaussien centré de covariance C_{AA}. De surcroit, on peut établir que le passage de de Y^c à Y^n est *réversible* dès qu'il est de même du passage de Y^c_A à Y^n_A pour tout bloc pivot A.

Exemples

Pour illustrer cet algorithme propagatif, nous montrons la simulation d'un vecteur gaussien sur une grille de 100×100 à maille unité. Sa matrice de covariance C est construite à partir de la fonction hyperbolique

$$C(h) = \left(1 + \frac{|h|}{20}\right)^{-1}$$

Plutôt que de générer un pivot à chaque itération, on procède par balayages aléatoires (scans), au cours desquels les noeuds de la grille sont ordonnés au hasard pour servir

de pivot une et une seule fois. La mise à jour de la valeur de chaque pivot se fait indépendamment de sa valeur courante. Partant d'un état initial identiquement nul, la figure 1 montre l'évolution du variogramme de la simulation en fonction du nombre de balayages.

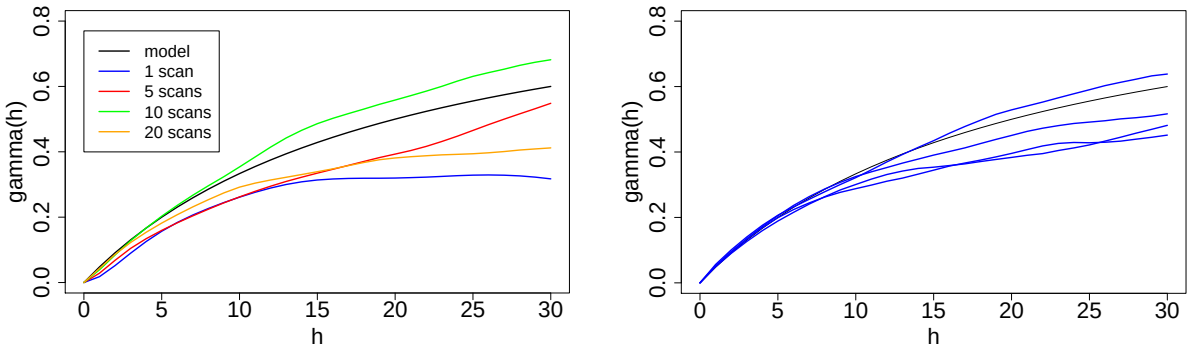


Figure 1: A gauche, variogrammes de simulations de covariance hyperbolique en fonction du nombre de balayages. A droite les variogrammes de 5 simulations obtenus avec 100 balayages

Une telle opération peut être répétée. Il s'avère qu'en moins de 100 balayages les fluctuations statistiques obtenues sont compatibles avec celles que l'on est en droit d'attendre du modèle. La figure 1 montre aussi les variogrammes expérimentaux de 5 simulations indépendantes, obtenues en 100 balayages.

D'autres essais ont été effectués avec d'autres modèles de covariance, en particulier les covariances dites stables

$$C_\alpha(h) = \exp\left(-\left(\frac{|h|}{a}\right)^\alpha\right) \qquad 0 < \alpha \leq 2$$

Les résultats obtenus avec le facteur d'échelle $a = 30$ sont reproduits en figure 2. Pour $\alpha = 0.5$, la convergence est très rapide; moins de 50 balayages suffisent à restituer correctement les fluctuations statistiques du variogramme. Ce nombre augmente en fonction de α. De l'ordre de la centaine dans le cas $\alpha = 1$ (comme pour la covariance hyperbolique), il passe à 150 pour $\alpha = 1.5$. Lorsque $\alpha = 2$, la convergence de l'algorithme n'est pas atteinte au bout de 200 balayages.
Comment interpréter ces résultats? Les quatre covariances considérées diffèrent essentiellement par leur comportement à l'origine, non différentiable de pente infinie pour $\alpha = 0.5$, non différentiable de pente finie pour $\alpha = 1$, une fois différentiable pour $\alpha = 1.5$ et infiniment différentiable pour $\alpha = 2$. Il semble donc que la vitesse de convergence dépend en bonne partie du degré de régularité de la covariance au voisinage de l'origine.

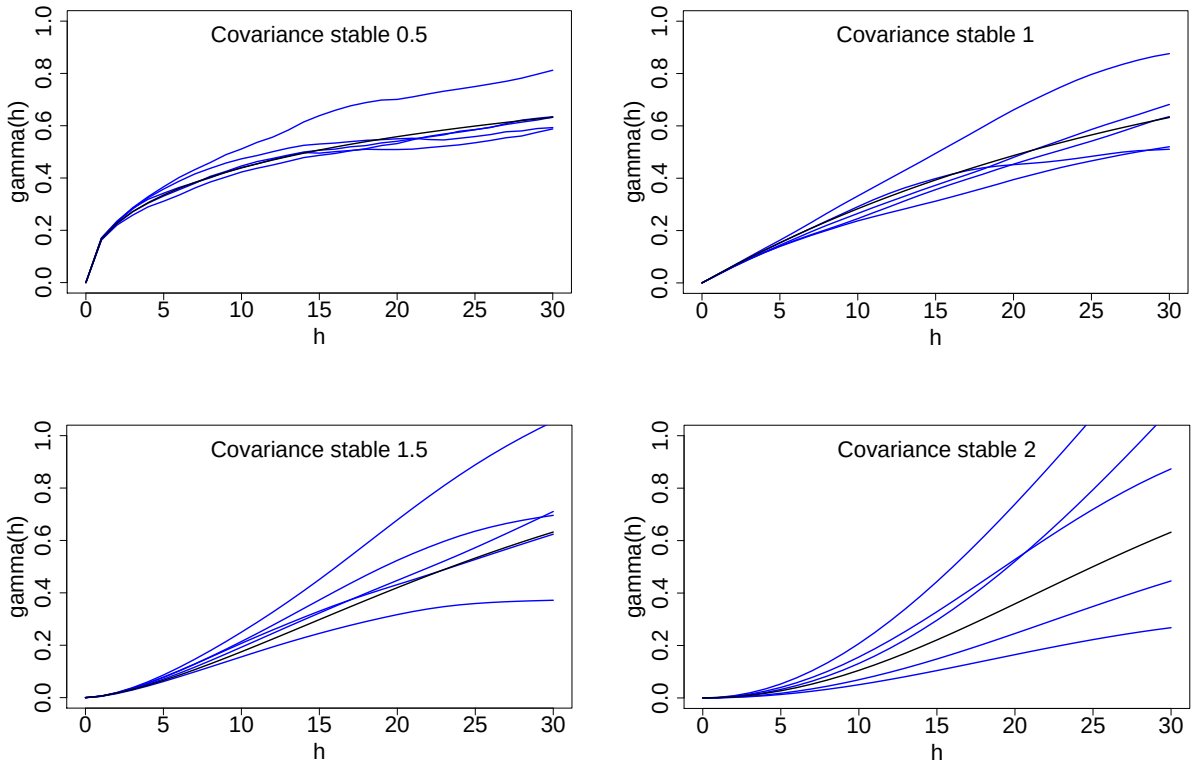


Figure 2: Fluctuations statistiques associées aux covariances stables d'ordre 0.5, 1 (exponentiel), 1.5 et 2 (gaussien) après 50, 100, 150 et 200 balayages. L'algorithme de simulation converge d'autant plus vite que la covariance est moins régulière au voisinage de l'origine.

Il est à noter qu'une stratégie de propagation à partir de pivots non ponctuels s'avère bien plus efficace. Dans le cas de la covariance gaussienne ($\alpha = 2$), la figure 3 montre que la convergence s'obtient en 50 balayages avec des pivots de 5 pixels. L'algorithme marche de la façon suivante: à chaque balayage, les 10000 pixels du champ sont ordonnés au hasard et regroupés par paquets de 5 pour servir de pivots. On dispose ainsi de $10000/5 = 2000$ pivots, c'est-à-dire 2000 itérations par balayage.

Quelques mots sur la vitesse de convergence de l'algorithme

Lorsque la simulation est initialisée par la réalisation d'un vecteur gaussien de matrice de covariance $C^{(0)}$, le vecteur obtenu au terme du n^{ème} balayage par mise à jour indépendante est aussi gaussien. Sa matrice de covariance $C^{(n)}$ est liée à $C^{(0)}$ et C par la relation (Desassis, 2011)

$$C^{(n)} - C = B^n (C^{(0)} - C)^t B^n. \tag{2}$$

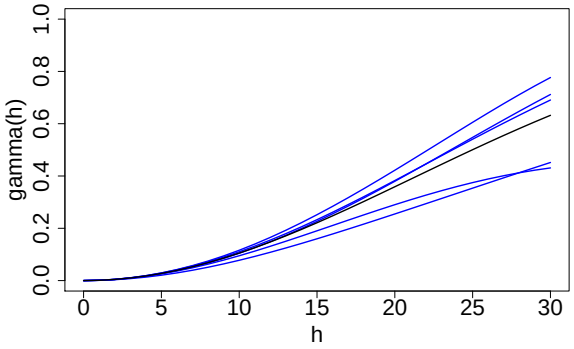


Figure 3: Fluctuations statistiques associées à la covariance gaussienne avec des pivots de 5 composantes avec 50 balayages

Il s'ensuit que la vitesse de convergence de l'algorithme est régie par le carré du rayon spectral de B, d'où l'intérêt de son expression. On peut montrer que l'on a

$$B = (Id - L)^{-1}U,$$

où Id, L et U sont respectivement les matrices diagonale (identité), triangulaire inférieure et triangulaire supérieure de la matrice C.

Plus généralement, lorsque $y_\alpha^n = ry_\alpha^c + \sqrt{1-r^2}u$, $C^{(n)}$ vérifie une équation similaire à (2), mais B est remplacée par la matrice

$$B_r = [Id - (1-r)L]^{-1}\,[rId + (1-r)U]$$

Le théorème d'Ostrowsky-Reich implique que le rayon spectral de B_r est strictement inférieur à 1, ce qui assure que la vitesse de convergence de l'algorithme est géométrique.

On peut être amené à penser que cette vitesse va se dégrader au fur et à mesure que r s'éloigne de 0. Rien n'indique qu'il en soit ainsi. On constate expérimentalement que le rayon spectral de B_r n'atteint pas son minimum en $r = 0$, mais plutôt pour des valeurs de r négatives, souvent entre -0.3 et -0.6.

Bibliographie

Desassis, N. (2011) Note sur l'algorithme de propagation de Gibbs. Rapport technique, MinesParistech.

Galli, A. et Gao, H. (2001) Rate of convergence of the Gibbs sampler in the Gaussian cas. *Mathematical Geology*, 33-6, 653–677.

Geman, S. et Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6-6, 721–741.

Lantuéjoul, C. (2011) Three dimensional conditional simulation of a Cox process using multisupport samples: methodological aspects. Rapport technique, MinesParistech.

ESTIMATION FOR RANDOM COEFFICIENT AUTOREGRESSIVE MODELS ON A PLANE

Kharfouchi Soumia* and Mehri Houd**
*Département de Mathématique, Université Mentouri Constantine
s_kharfouchi@yahoo.fr

**Research Unit in Logistics, Industrial Management and Quality (LOGIQ) - ISGI-Sfax. BP n°954-3018 Sfax-TUNISIA

mehri.houda@gmail.com

Abstract: In this paper, we introduce a new class of non linear spatial models. This class is generated by a spatial random coefficient autoregressive process. A sufficient condition for the existence of a stationary and ergodic solution is given and a quasi-maximum likelihood estimation procedure is proposed.
***Keywords*:** Random Coefficient Autoregressive model. Spatial statistics. quasi-maximum likelihood.

1 Introduction

The study of random coefficient autoregressive processes (RCA) a long a line (time t) has received much attention in the economic literature: see, for example, Kendal (1953). These processes describe the fact that a random variable X_t arises as the linear combination of p preceding variables $a_1(t)X_{t-1} + \cdots + a_p(t)X_{t-p}$ plus a random deviation, where $a_1(t), \dots, a_p(t)$ are random variables. Relatively few results have yet emerged for comparable processes on a plane. Unlike the time series RCA case, several kind of spatial random coefficient autoregressive models (SRCA) may be defined, most of the different representations appear to depend on the order chosen on the lattice $\mathbb{Z}^2$. These models incorporate spatial effects, they can be applied in regional science, labor economics, and real estate economics. Hence, our model is a $\mathbb{R}$-valued spatial process on a regular rectangular grid defined in two dimensions with sites labeled (i,j), with an associate random variable $X(i,j)$ defined at each site, as

$$X(i,j) = \sum_{(k_1,k_2)\in S]\mathbf{0},\boldsymbol{p}]} \left[\beta_{k_1k_2} + \beta_{k_1k_2}(i,j)\right] X(i-k_1, j-k_2) + \varepsilon(i,j), \qquad (1.1)$$

where the index set $S]\mathbf{0},\boldsymbol{p}]$ is defined as

$$S]\mathbf{0},\boldsymbol{p}] = \{(l,m)\in\mathbb{Z}^2, (l,m) \prec \boldsymbol{p} \text{ or } (l,m) = \boldsymbol{p}\} - \{(0,0)\},$$

and $\prec$ denote the lexicographic order on $\mathbb{Z}^2$(i.e. for all $\boldsymbol{s} = (s_1, s_2)$ and $\boldsymbol{t} = (t_1, t_2)$ in $\mathbb{Z}^2$, $\boldsymbol{s} \prec \boldsymbol{t}$ ($\boldsymbol{s}$ is prior to $\boldsymbol{t}$) if and only if $[(s_1 < t_1)\ or\ (s_1 = t_1 \text{ and } s_2 < t_2)]$).
for this model we need the following assumptions

i) $\{\varepsilon(i,j), (i,j)\in\mathbb{Z}^2\}$ is an i.i.d. sequence of random variables with mean zero and variance σ^2.

ii) The $\beta_{k_1k_2}, k_1 = 1, \dots, p_1, k_2 = 1, \dots, p_2$, are real constants.

iii) $\left(\beta_{k_1k_2}(i,j),\ (i,j)\in\mathbb{Z}^2\right)_{\substack{1\le k_1\le p_1\\ 1\le k_2\le p_2}}$ is a sequence of i.i.d. random variables with zero mean. $\left(\beta_{k_1k_2}(i,j)\right)_{\substack{1\le k_1\le p_1\\ 1\le k_2\le p_2}}$ is independent of $\{\varepsilon(i,j)\}$ for all $(i,j)\in\mathbb{Z}^2$.

iv) Letting

$$\mathcal{B}_1(i,j)=\begin{pmatrix}\underline{\beta}_0(i,j) & \cdots & \underline{\beta}_{p_1}(i,j)\\ 0 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \ldots & 0\end{pmatrix},\ \mathcal{B}_2(i,j)=\begin{pmatrix}\bar{\beta}_1(i,j) & \cdots & \bar{\beta}_{p_1}(i,j) & \bar{0}\\ 0 & \cdots & 0 & \bar{0}\\ \vdots & & \vdots & \vdots\\ 0 & \ldots & 0 & \bar{0}\end{pmatrix}$$

where $\underline{\beta}_k(i,j)$, $\bar{\beta}_s(i,j)$ and $\bar{0}$ are p_2-vectors such that,

$\underline{\beta}_k(i,j)=\left(\beta_{k1}(i,j),\cdots,\beta_{kp_2}(i,j)\right)$, $\bar{\beta}_k(i,j)=\left(0,\cdots,0,\beta_{s0}(i,j)\right)$, $k=0,\cdots,p_1$ and $\bar{0}=(0,\cdots,0)$, we suppose that

$$E\{\mathcal{B}_1(i,j)\otimes\mathcal{B}_1(i,j)\}=\mathfrak{C}_1 \text{ and } E\{\mathcal{B}_2(i,j)\otimes\mathcal{B}_2(i,j)\}=\mathfrak{C}_2.$$

To establish the existence of a stationary and ergodic solution of equation (1.1) we need to represent the ***p***th order SRCA model as a state-space version of equation (1.1), obtained by defining the $(p_1+1)p_2\times 1$ vectors

$$\underline{X}(i,j)=\left(x_1(i,j),\cdots,x_{p_1}(i,j)\right)' \text{ and } \underline{\varepsilon}(i,j)=\left(0,\cdots,0,\varepsilon(i,j)\right)$$

where $x_k(i,j)=\left(X(i-k,j),X(i-k,j-1),\cdots,X(i-k,j-p_2+1)\right)$,

we also need the $(p_1+1)p_2\times(p_1+1)p_2$ matrices

$$\mathcal{M}_1=\begin{pmatrix}\underline{\beta}_0 & \cdots & \underline{\beta}_{p_1}\\ 0 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \ldots & 0\end{pmatrix},\mathcal{M}_2=\begin{pmatrix}\bar{\beta}_1 & \cdots & \bar{\beta}_{p_1} & \bar{0}\\ 0 & \cdots & 0 & \bar{0}\\ \vdots & & \vdots & \vdots\\ 0 & \ldots & 0 & \bar{0}\end{pmatrix}$$

where $\underline{\beta}_k=\left(\beta_{k1},\cdots,\beta_{kp_2}\right)$, $\bar{\beta}_k=(0,\cdots,0,\beta_{s0})$, $k=0,\cdots,p_1$.

With this notation, equation (1.1) can be rewritten in the form

$$\underline{X}(i,j)=[\mathcal{M}_1+\mathcal{B}_1(i,j)]\underline{X}(i,j-1)+[\mathcal{M}_2+\mathcal{B}_2(i,j)]\underline{X}(i-1,j)+\underline{\varepsilon}(i,j), \qquad (1.2)$$

.

2 Conditions for stationarity

In this section we shall investigate the conditions under which the process (1.2) is strictly stationary and ergodic. Our strategy largely follows the time domain approach of Lui and Brockwel (1990) in the case of general bilinear time series model.

.

3 GMM estimation

In this section, a quasi-maximum likelihood estimation procedure is proposed for SRCA models. The estimators are shown to be consistent and asymptotically normal

REFERENCES

Anselin, L. 1988. *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic Publishers.

Kelejian, H. H. and Prucha, I. 1999. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model *International Economic Review* **40**: 509-533.

Kendall, M. G. 1953. The analysis of economic time series. Part I: Prices. *J. Roy. Statist. Soc*, A **96**, 11-25.

Liu, J. and Brockwell, P. J.1988. On the general bilinear time series model, *J. Appl. Prob*. **25**, 553-564.

A factor model approach for the segmentation of correlated time series

Emilie Lebarbier & Stéphane Robin

UMR 518 AgroParisTech / INRA Math. Info. Appli.
16, rue Claude Bernard
75005 Paris France

Résumé. On s'intéresse à la segmentation d'un ensemble de séries d'observations corrélées, typiquement organisées spatialement. On propose de modéliser la dépendance entre les séries au moyen d'un modèle à facteur. Cette modélisation de la dépendance autorise l'utilisation d'algorithmes de segmentation efficaces pour obtenir les estimateur du maximum de vraisemblance. On propose également une procédure de sélection de modèle pour déterminer le nombre de points de ruptures ainsi que le nombre de facteurs.

Abstract. We consider the segmentation of set of correlated time-series, typically with some spatial structure. We propose to model the between-series dependency with a factor model. This modelling allows us to use efficient segmentation algorithms to obtain the maximum likelihood estimates. We also propose a model selection procedure to determine the number of change point in each series and the number of factors.

Keywords. E-M algorithm, Factor model, Segmentation.

1 Introduction

We consider the detection of change-points in a set of time-series. We are typically interested in the case where this data consist in series of measurements observed along time in different locations. Each series is supposed to be affected changes at series-specific breakpoints and the signals observed at each location are supposed to be correlated due to their spatial organisation.

One of the difficulty here is to propose a modelling that leads to an efficient estimation algorithm. Indeed, the inference of segmentation models often require to search over the space of all possible segmentations, which is prohibitive in terms of computational time. Dynamic programming (DP) strategies remain among the most efficient but can only be applied when the contrast to be optimised is additive with respect to the segments. In presence of dependency, the contrast (e.g. the log-likelihood) is generally not additive. Our strategy consists in 'removing' the dependency so that, at a given step of the estimation algorithm, dynamic programming can be applied to transformed data.

A similar setting is considered in [2] where a variance component model is used to account for the dependecy between the series. Our purpose here is broaden the set of possible dependency structures that the modelling can account for. The factor model provides a convenient and efficient way to describe various covariance matrices, still limiting the number of parameters. It can be viewed as a generalisation of variance components models, where the components are unknown and need to be estimated, together with the associated variances. It is based on the spectral decomposition of the covariance matrix and has been successfully applied in situation where very little is known about the correlation structure (see e.g. [1]). The inference of factor models is often achieved via an EM algorithm.

We present here a general model for correlated Gaussian time series, based on a factor model for the covariance matrix. We show that some by-product of the EM algorithm can be used to 'remove' the dependency between the series. This allows use to combine EM and DP algorithms together. We then discuss the issue of choosing both the number of breakpoints and the number of factors. Simulation and illustrations will be presented in the oral presentation.

2 Data, notations and model

Data and notations. We consider M series with n_m points each. We note y_{tm} the observed signal of series m at time t. The total number of observations is $N = \sum_{m=1}^{M} n_m$. In the following, we consider that $n_m = n$ whatever the series. The data y are modeled by a random Gaussian process Y with size $[n \times M]$. In general we denote A_t the row vector with size M of the matrix A and A^m its column vector with size n. Thus Y^m represents whole series m, while Y_t stands for the observations at time t in all the series.

Segmentation. We consider here that each series has its own segmentation: the mean of the series $\{Y_{tm}\}_t$ is subject to $K_m - 1$ specific abrupt changes at breakpoints $\{t_k^m\}$ (with convention $t_0^m = 0$ and $t_{K_m}^m = n_{\max}$) and is constant between two breakpoints within the interval $I_k^m =]t_{k-1}^m, t_k^m]$. In the following we denote by $K = \sum_m^M K_m$ the total number of segments and $n_k^m = t_k^m - t_{k-1}^m$ the length of segment k for series m ($k = 1, \ldots, K_m$). The segmentation model is written as follows:

$$Y_{tm} = \mu_{km} + F_{tm} \qquad \forall t \in I_k^m \tag{1}$$

where the error vectors $\{F_t\}_t$ has a centered Gaussian distribution with a covariance matrix $\mathbf{\Sigma}$ to be specified.

Correlations between series. We want to take into account the correlations that can exit between series. In the case of spatial data, the correlation can take the form $e^{-d(m,m')}$ where $d(m, m')$ is the distance between the locations of series m and m'. However this

correlation structure (or other structures) hampers the use of the DP algorithm to obtain the best segmentation.
We propose here to consider the factor analysis model framework. That consists in identifying a linear space of Q random vectors that captures the dependence among the series (see [1]). In others words, we set

$$\mathbf{\Sigma} = \mathbf{B}\mathbb{V}(\mathbf{Z})\mathbf{B}' + \mathbf{\Psi},$$

where $\mathbf{Z}$ corresponds to the Q random vectors and B the associated coefficients. The dependence between series is then free from any spatial structure. In the above decomposition, $\mathbf{B}\mathbb{V}(\mathbf{Z})\mathbf{B}'$ refers to the shared variance and $\mathbf{\Psi}$ to the specific one.

Model. With the previous decomposition of the variability, the model (1) can be rewritten as a mixed linear model:

$$Y_{tm} = \mu_{km} + \sum_{q=1}^{M} Z_{tq} b_{qm} + E_{tm} \qquad \forall t \in I_k^m$$

where the $\{E_t\}_t$ are i.i.d. centered Gaussian vectors with variance matrix $\mathbf{\Psi}$ and the $\{Z_t\}_t$ are i.i.d. centered Gaussian vectors with variance I_Q (without loss of generality), the two sets of vectors being independent. So $\mathbf{\Sigma}$ can be decomposed as

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi}. \tag{2}$$

Here $\mathbf{\Psi}$ is supposed to be diagonal $\mathbf{\Psi} = \sigma^2 I_M$. This will allow us to use the DP algorithm for the segmentation parameter estimation. The matrix formulation of this linear model is

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{B}' + \mathbf{E} \tag{3}$$

where

- $\mathbf{Y}$, with size $[n \times M]$, stands for the observed data,
- $\mathbf{T}$ is the incidence matrix of breakpoints with size $[n \times K]$, $T^m = \text{Bloc}\left[\mathbb{1}_{n_{K_m}^m}\right]$, and $\mathbf{T} = \left[T^1\ T^2\ \ldots\ T^M\right]$.
- $\boldsymbol{\mu}$ the means with size $[K \times M]$ (and μ_k^m the mean of the segment k for series m) such that $\mu^m = \text{Bloc}\left[\mu_{K_m}^m\right]$, and $\boldsymbol{\mu} = \left[\mu^1\ \mu^2\ \ldots\ \mu^M\right]$.
- $\mathbf{Z}$ with size $[n \times Q]$ and $\mathbf{B}$ with size $[M \times Q]$,
- $\mathbf{E}$ with size $[n \times M]$.

The main difference between this model and a classical mixed linear model is that both the incidence matrix $\mathbf{T}$ and the factor matrix $\mathbf{B}$ are unknown.

3 Estimation using the EM algorithm

We now consider the maximum-likelihood inference of model (3). This can be done via an EM algorithm, considering that $\mathbf{Z}$ represents the missing data (hidden and unobserved). The parameters of the model are $\phi = (\mathbf{T}, \boldsymbol{\mu}, \sigma^2, \mathbf{B})$. In this setting, the complete-data log-likelihood is:

$$\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi) = \log \mathcal{L}_0(\mathbf{Y}|\mathbf{Z}; \phi) + \log \mathcal{L}_1(\mathbf{Z}).$$

Since the distribution of $\mathbf{Z}$ does not depend on the parameters ϕ, only the first term will be considered which is written:

$$-2 \log \mathcal{L}_0(\mathbf{Y}|\mathbf{Z}; \phi) = N \log(2\pi) + n \log(|\boldsymbol{\Psi}|) + \sum_{t=1}^{n} \|Y_t - \mu_{k(t)} - Z_t \mathbf{B}'\|^2_{\boldsymbol{\Psi}^{-1}},$$

then its conditional expectation, $Q_0(\phi; \phi^{(h)}) = \mathbb{E}_{\phi^{(h)}} \{\log \mathcal{L}_0(\mathbf{Y}|\mathbf{Z}; \phi)|\mathbf{Y}\}$ satisfies

$$-2Q_0(\phi; \phi^{(h)}) \quad = \quad N \log(2\pi) + n \log(|\boldsymbol{\Psi}|) + \sum_{t=1}^{n} \left[\|Y_t - \mu_{k(t)} - \widehat{\mathbf{Z}}_t^{(h)} \mathbf{B}'\|^2_{\boldsymbol{\Psi}^{-1}} + \mathrm{Tr}\left(\mathbf{B}' \boldsymbol{\Psi}^{-1} \mathbf{B} \mathbf{W}_t^{(h)} \right) \right],$$

where $\mathbb{E}_\phi\{\cdot\}$ is the expectation operator using ϕ as the parameter value and $\mathbb{V}_\phi\{\cdot\}$ the corresponding variance, $\widehat{\mathbf{Z}}_t^{(h)} = \mathbb{E}_{\phi^{(h)}} \{\mathbf{Z}_t|\mathbf{Y}\}$, $\mathrm{Tr}(A)$ is the trace of matrix A, $|A|$ its determinant and $\mathbf{W}_t^{(h)} = \mathbb{V}_{\phi^{(h)}} \{\mathbf{Z}_t|\mathbf{Y}\}$.

E-step This step consists in the calculation of the conditional expectation which only requires the calculation of $\widehat{\mathbf{Z}}$ and $\mathbf{W}$: at iteration $(h+1)$, we get

$$\begin{cases} \widehat{\mathbf{Z}}_t^{(h+1)} = \tilde{Y}_t^{(h)} \mathbf{B}^{(h)} \mathbf{W}_t^{(h)} / \sigma^{2,(h)}, \\ \mathbf{W}_t^{(h+1)} = \left(I_M + \mathbf{B}'^{(h)} \mathbf{B}^{(h)} / \sigma^{2,(h)} \right)^{-1}. \end{cases}$$

where $\tilde{Y}_t^{(h)} = Y_t - \mu_{k(t)}^{(h)}$.

M-step This step consists in the estimation of the parameters by maximizing the obtained conditional expectation.

- Estimation of the variance component σ^2:

$$\sigma^{2,(h+1)} \quad = \quad \frac{1}{N} \sum_{t=1}^{n} \left[E_t^{(h)} E_t'^{(h)} + \mathrm{Tr}\left(\mathbf{B}'^{(h)} \mathbf{B}^{(h)} \mathbf{W}_t^{(h+1)} \right) \right],$$

  where $E_t^{(h)} = Y_t - \mu_{k(t)}^{(h)} - \widehat{\mathbf{Z}}_t^{(h)} \mathbf{B}'^{(h)}$.

- Estimation of $\mathbf{B}$.

$$\mathbf{B}^{(h+1)} = \sum_{t=1}^{n} (Y_t - \mu_{k(t)}^{(h)})' \widehat{\mathbf{Z}}_t^{(h+1)} \left[\sum_{t=1}^{n} (\widehat{Z}_t'^{(h+1)} \widehat{Z}_t^{(h+1)} + W_t^{(h+1)}) \right]^{-1}.$$

- Estimation of the segmentation parameters $\mathbf{T}\boldsymbol{\mu}$.

$$\begin{aligned} \left\{ \mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)} \right\} &= \arg\max_{\mathbf{T},\boldsymbol{\mu}} Q_0, \\ &= \arg\max_{\mathbf{T},\boldsymbol{\mu}} \sum_{t=1}^{n} \| Y_t - (\mathbf{T}\boldsymbol{\mu})_t - \widehat{\mathbf{Z}}_t^{(h+1)} \mathbf{B}'^{(h+1)} \|^2_{\boldsymbol{\Psi}^{-1},(h+1)}, \\ &= \arg\max_{\mathbf{T},\boldsymbol{\mu}} \sum_{t=1}^{n} \| \tilde{\tilde{Y}}_t - (\mathbf{T}\boldsymbol{\mu})_t \|^2_{\boldsymbol{\Psi}^{-1},(h+1)}. \end{aligned}$$

  where $\tilde{\tilde{Y}}_t = Y_t - \widehat{\mathbf{Z}}_t^{(h+1)} \mathbf{B}'^{(h+1)}$. This last term can be viewed as a correction to remove dependency between the series. Since $\boldsymbol{\Psi}$ is diagonal the DP (in particular the two-stages DP) can be used to obtain the segmentation parameter estimations.

4 Model selection

Here either the number of factors q and the number of segments K should be estimated. The joint estimation of two parameters is not classical. We propose here a heuristic to select both these two parameters. First, for each K, q is selected by using the BIC criterion:

$$\begin{aligned} \hat{q}_K &= \operatorname*{argmin}_{q \in \{1,\ldots,M-1\}} BIC(q_K), \\ &= \operatorname*{argmin}_{q \in \{1,\ldots,M-1\}} -2\mathcal{L}(\widehat{\mathbf{T}\boldsymbol{\mu}}_K, \hat{\boldsymbol{\Sigma}}_q) + Dq \log(n), \end{aligned}$$

where $\mathcal{L}(\widehat{\mathbf{T}\boldsymbol{\mu}}_K, \hat{\boldsymbol{\Sigma}}_q)$ is the log-likelihood calculated at its maximum for a fixed K and a fixed q, and D_q is the number of parameters in a model with q factors, $D_q = q(2M - q + 1)/2 + 1$. Indeed for the variance components, according to the variance decomposition (cf equation (2)), the number of parameters are $M \times q$ for B and one for σ^2. Moreover, with the orthogonality condition on B, only $Mq - q(q-1)/2$ need to be estimated.
To select the number of segments, we use the modified BIC proposed by [3] adapted to the joint segmentation by [2]

$$\begin{aligned} mBIC_{\text{JointSeg}}(K) &= \left(\frac{N+1}{2} \right) \log SS_{\text{all}} - \left(\frac{N-K+1}{2} \right) \log SS_{\text{wg}}(\hat{t}) + \log \left[\Gamma \left(\frac{N-K+1}{2} \right) \right] \\ &\quad - \frac{1}{2} \sum_{m=1}^{M} \sum_{k=1}^{k_m} \log \hat{n}_k^m + \left(\frac{1}{2} - (K - M) \right) \log(N), \end{aligned}$$

where

$$SS_{\text{wg}}(\hat{t}) = \sum_{t=1}^{n} (\mathbf{Y}_t - \hat{\mu}_{k(t)}) \hat{\mathbf{\Sigma}}_q^{-1} (\mathbf{Y}_t - \hat{\mu}_{k(t)})',$$

$$SS_{\text{all}} = \sum_{t=1}^{n} (\mathbf{Y}_t - \bar{Y}) \hat{\mathbf{\Sigma}}_q^{-1} (\mathbf{Y}_t - \bar{Y})',$$

with $\hat{n}_k^m$ is the length of segment k in profile m ($\hat{n}_k^m = \hat{t}_k^m - \hat{t}_{k-1}^m + 1$), and $\hat{mu}_{k(t)}$ is a vector of size M with the component m is $\bar{y}_{mk} = (\hat{n}_k^m)^{-1} \sum_{t=\hat{t}_{k-1}^m+1}^{\hat{t}_k^m} y_m(t)$ if $t \in \hat{I}_k^m$.

References

[1] C. Friguet, M. Kloareg, and D. Causeur, *A factor model approach to multiple testing under dependence*, J. Amer. Statist. Assoc. **104** (2009), no. 488, 1406–15, DOI:10.1198/jasa.2009.tm08332.

[2] F. Picard, E. Lebarbier, E. Budinska, and S. Robin, *Joint segmentation of multivariate gaussian processes using mixed linear models*, Comput. Statist. and Data Analysis **55** (2011), no. 2, 1160–70.

[3] N. R. Zhang and D. O. Siegmund, *A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data*, Biometrics **63** (2007), no. 1, 22–32.

Fiabilité - Qualité

Estimation du taux de défaillance pour des équipements industriels sous contraintes d'environnement, *Lise Guérineau and Evans Gouno*

Certains équipements industriels évoluent dans des conditions environnementales fluctuantes qui agissent sur leur niveau de fiabilité. Nous proposons une modélisation du taux de défaillance intégrant l'effet de l'environnement. Cette modélisation a été motivée par un cas concret concernant des composants du réseau électrique, pour lesquels les pannes apparues sur le parc sont répertoriées. L'approche s'appuie donc sur l'observation des matériels en conditions réelles d'utilisation.

Le taux de défaillance est supposé constant par morceaux et la valeur de celui-ci sur chaque intervalle de temps est régie par un même modèle combinant deux effets :

- l'effet de la température extérieure, intégrée dans le modèle par un facteur d'accélération de type Arrhenius,
- l'effet d'un phénomène de diffusion d'humidité, intrinsèque à l'équipement, intégré dans le modèle par un facteur d'accélération de type Peck.

Ce genre de modélisation a déjà été développé dans le cadre des essais accélérés mais dans ce contexte, le taux de défaillance est supposé monotone et les contraintes sont contrôlées. Dans notre situation, le taux n'est plus monotone et les contraintes ne sont pas contrôlées. Cependant, elles sont connues. On a un taux de défaillance dit en montagnes russes.

On utilise alors une méthode du maximum de vraisemblance pour estimer les paramètres du modèle. La démarche est éprouvée sur des données simulées, puis appliquée sur le jeu de données réelles qui a motivé l'étude.

Les résultats permettent aux industriels d'améliorer leurs connaissances sur la longévité et la vulnérabilité de leur matériel face à des agressions liées à l'environnement.

Détection de la défaillance des entreprises tunisiennes par la régression logistique semi paramétrique et les réseaux de neurones, *Sami Mestiri and Abdeljelil Farhat*

L'objectif de cet article est de comparer deux techniques de classification des entreprises : la régression logistique semi paramétrique et les réseaux de neurones dans le but de prévoir le risque de crédit des banques tunisiennes. L'échantillon utilisé comporte 528 firmes tunisiennes de différents secteurs d'activités dont nous disposons des bilans et des comptes financiers des exercices 1999-2006. Nous avons démontré que l'examen soigneux du rapport fonctionnel entre les ratios et la probabilité de la détresse et l'emploi de modèles basés sur les réseaux de neurones améliore la qualité des prévisions des modèles en terme de bon classement.

Fuzzy multivariate cumulative sum and exponentially weighted moving average control chart, *Ali Achouri and Hassen Taleb*

In this article, fuzzy set theory is proposed to construct multivariate cumulative sum (cusum)& exponentially weighted moving average (ewma)control chart. Real data taken from "Id´eal Sanitaire" process is used.Product units are classified into 5 categories. Each sample is presented by a single fuzzy set and each category is described by a linguistic variable. The resulted control charts are compared and analysed.

Multi-scale process monitoring in Nanomanufacturing, *Sihem Ben Zakour and Hassen Taleb*

The detection of the end of polishing during the chemical mechanical planarization (CMP)process is a critical task in semiconductor manufacturing. The disadvantages of offline approach has incited the researchers to discover an efficient substitute. In this paper, an alternative approach has presented named online method in which a sequential probability ratio test (SPRT) was developed and applied to the wavelet decomposed Acoustic emission data collected during the progression of the CMP process. This test is shown to be efficient in controlling complex processes and appropriated for real-time application by developing a moving block strategy.

A Data Depth Based EWMA Control Chart, *Amor Messaoud, Giovanni Porzio, Hela Abidi and Mohamed Limam*

Nonparametric control charts are attractive in many industrial applications because they do not require knowledge about the shape of the underlying distribution. A new nonparametric control chart for multivariate processes is proposed. It is based on the concept of data depth approach. Its design is discussed and it is applied to a data set from a drilling process.

Estimation du taux de défaillance pour des équipements industriels sous contraintes d'environnement

Lise Guérineau[(a),(b)] & Evans Gouno[(b)]

(a) *EDF R&D*
1, avenue du Général de Gaulle
BP 408 – 92141 CLAMART

(b) *Université de Bretagne Sud*
Campus de Tohannic
BP573 – 56017 VANNES

Mots-clés : Taux de défaillance par morceaux, Maximum de vraisemblance, Censure, Modèle d'accélération, Peck, Arrhenius

Résumé
Cet article se propose de traiter des données de fiabilité concernant des matériels exposés à un environnement fluctuant. Une méthode est envisagée pour estimer le taux de défaillance ainsi que des paramètres rendant compte de l'effet des conditions physiques (température, diffusion d'humidité). La méthode est éprouvée sur des données simulées, puis appliquée sur un jeu de données réelles.

Key words : Piecewise exponential failure rate, Maximum Likelihood, Censoring, Acceleration model, Peck, Arrhenius

Abstract
This article deals with reliability data for systems under time-varying environmental conditions. A method is proposed to estimate the failure rate, incorporating parameters that relate to the effect of physical conditions, ie temperature and moisture diffusion. The method is assessed on simulated data and is applied on a real data set.

Introduction

Carer et al. (2010) proposent d'utiliser un modèle de type Cox pour intégrer dans l'étude de la fiabilité de matériels électriques des contraintes physiques. Le travail présenté ici reprend cette idée et développe une méthode d'inférence pour analyser des données concernant des équipements mis en service à différentes dates et évoluant sous des profils de contraintes qui varient dans le temps. L'objectif est d'étudier la relation entre le taux de défaillance et l'environnement. En particulier, on s'intéresse à l'effet de la température extérieure et à un phénomène intrinsèque à l'élément, phénomène de diffusion d'humidité. La modélisation de ce dernier s'appuie sur des résultats proposés par des auteurs tels que Crank (1979) et Tencer (1994). Contrairement à ce que supposent de nombreuses études, la fiabilité dans ce contexte n'est pas nécessairement monotone. On a en général un taux de défaillance dit *en montagnes russes*.

L'estimation de taux de défaillance possédant cette propriété a été étudiée par Kim et Proschan (1991). Nous envisageons ici d'intégrer dans le modèle de Kim et Proschan, des modèles d'accélération de type Arrhenius / Peck. L'approche est voisine des méthodes proposées par Gouno (2007) dans le cadre des essais accélérés en step-stress. Le retour d'expérience sur les pannes peut être vu comme le résultat d'un test « grandeur nature » où les contraintes environnementales ne sont pas contrôlées. Des modèles de fiabilité basés sur des observations en conditions réelles ont déjà été étudiés par Monroe et Pan (2008). Notre approche diffère dans le sens où aucune distribution n'est ajustée aux variables d'environnement : elles sont soit observées, soit modélisées de manière à approximer au mieux la réalité.

1 Modèle et Hypothèses

Dans une période de temps donnée, découpée en m intervalles de longueur τ, des équipements sont mis en service à différentes dates (dates de pose). On observe des dates de panne pour certains éléments. A un élément posé dans un intervalle quelconque, on associe une date de pose correspondant à la borne inférieure de cet intervalle. Ainsi, une date de pose est un multiple de τ. Les éléments évoluent dans des conditions environnementales dont le niveau varie avec le temps. Nous supposons cependant que celui-ci reste constant dans chaque intervalle.

Le modèle présenté est inspiré d'un cas concret où l'élément subit l'influence de la température, mais aussi de l'humidité qu'il contient. On note T_i, la valeur de la température dans le $i^{\text{ème}}$ intervalle. La valeur de l'humidité dépend du temps de service et sera fonction de la date de pose ; on note HR_{i-d}, sa valeur dans le $(i-d)^{\text{ème}}$ intervalle après sa date de pose d. On note T_{ref} la température moyenne et HR_{ref} l'humidité initiale.

On exprimera le taux de défaillance d'un élément mis en service à la date d par une fonction étagée de la forme :

$$\lambda_d(x) = \sum_{j=1}^{m-d} \lambda_{d+j} 1_{\left[(j-1)\tau;\ j\tau\right)}(x) \tag{1}$$

où pour alléger les écritures on a noté d pour d/τ. Cette hypothèse est équivalente à considérer que la durée de vie d'un élément dans le $j^{\text{ème}}$ intervalle après sa date de pose d, suit une loi exponentielle de paramètre λ_{d+j}. Le paramètre sera exprimé en fonction des conditions environnementales à l'aide d'un modèle d'accélération de type Peck/Arrhenius.

$$\lambda_{d+j} = \lambda_0 \exp\left\{-\frac{E_a}{K}\left(\frac{1}{T_{d+j}} - \frac{1}{T_{ref}}\right)\right\}\left(\frac{HR_j}{HR_{ref}}\right)^{\eta} = \lambda_0 \exp\{E_a y_{d+j} + \eta z_j\}$$

où
- λ_0 est le taux de référence,
- E_a l'énergie d'activation,
- $K = 8,61.10^{-5} eV/Kelvin$ la constante de Boltzmann et
- η un réel positif, le coefficient de Peck.

Ce type de modèle combine un facteur d'accélération lié à la température, décrit par Lall (1996) et un facteur d'accélération lié à l'humidité, introduit par Peck (1986).

2 Estimation

Les paramètres du modèle à estimer sont λ_0, E_a et η. Considérons l'élément i. Notons d_i sa date de pose. Sous l'hypothèse d'un taux de défaillance de la forme (1), sa durée de bon fonctionnement dans l'intervalle $[(\ell-1)\tau; \ell\tau[$ est une variable aléatoire dont la densité a pour expression :

$$f_{d_i}(x) = \lambda_{d_i+\ell} \exp\left\{ -\sum_{j=1}^{\ell-1} \lambda_{d_i+j}\tau - \lambda_{d_i+\ell}\big(x-(\ell-1)\tau\big) \right\}$$

On introduit δ_i l'indicateur de censure. $\delta_i = 1$ si la durée de bon fonctionnement de l'élément i est censurée à droite – c'est à dire si l'élément i ne présente pas de défaillance dans la période d'observation – et 0 sinon. Pour n éléments observés, la vraisemblance a donc pour expression :

$$L(\lambda_0, E_a, \eta) = \prod_{i=1}^{n} \left[f_{d_i}(x_i)\right]^{1-\delta_i} \left[R_{d_i}(m\tau - d_i\tau)\right]^{\delta_i}$$

où x_i est la durée de bon fonctionnement de l'élément i et

$$R_{d_i}(m\tau - d_i\tau) = \exp\left\{ -\int_0^{m\tau-d_i\tau} \lambda_{d_i}(s)ds \right\} = \exp\left\{ -\sum_{j=1}^{m-d_i} \lambda_{d_i+j}\tau \right\}$$

est la fonction de fiabilité pour un élément posé à la date d_i. On supposera que lorsqu'une défaillance survient dans un intervalle, elle a lieu au début de celui-ci. Ainsi, les x_i sont des multiples de τ. Comme pour les dates de pose, on allégera les écritures en notant x_i pour x_i/τ. La log-vraisemblance s'exprime alors par :

$$\begin{aligned}
\log L(\lambda_0, E_a, \eta) &= \sum_{i=1}^{n} \left((1-\delta_i)\log \lambda_{d_i+x_i+1} - \Big[(1-\delta_i) \sum_{j=1}^{x_i} \lambda_{d_i+j} + \delta_i \sum_{j=1}^{m-d_i} \lambda_{d_i+j} \Big]\tau \right) \\
&= k \log \lambda_0 + \sum_{i=1}^{n} (1-\delta_i)(E_a y_{d_i+x_i+1} + \eta z_{x_i+1}) - \lambda_0 \tau \sum_{i=1}^{n} U_i(E_a, \eta)
\end{aligned}$$

où $k = \sum_{i=1}^{n}(1-\delta_i)$ est le nombre de défaillances observées et

$$U_i(E_a, \eta) = \sum_{j=1}^{x_i}(1-\delta_i)\exp\{E_a y_{d_i+j} + \eta z_j\} + \sum_{j=1}^{m-d_i}\delta_i \exp\{E_a y_{d_i+j} + \eta z_j\}$$

Les estimateurs du maximum de vraisemblance de λ_0, E_a et η sont obtenus en utilisant un algorithme de Newton-Raphson.

3 Simulations et Applications

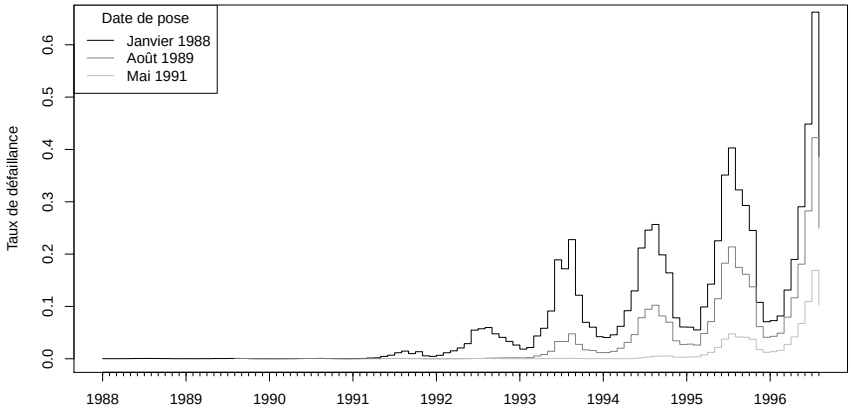
On génère des données avec différentes valeurs de paramètres pour $n = 50$ éléments suivis pendant 3 années civiles ($m = 36$ intervalles de longueur $\tau = 1$ mois). Le tableau 1 donne pour chaque jeu de paramètres la moyenne des estimations, l'écart-type ainsi que l'erreur quadratique moyenne.

Nous avons appliqué la méthode sur des données collectées sur un matériel électrique. La figure 1 donne une représentation du taux de défaillance estimé obtenu.

TABLE 1 – *Moyennes, biais et erreurs quadratiques moyennes (EQM) calculés à partir d'échantillons de 500 estimations, pour différentes combinaisons des paramètres (paramètres d'initialisation pour l'algorithme de Newton-Raphson :* $\lambda_0 = 0,1$, $E_a = 1$, $\eta = 3$*)*

	Exact	Moyenne	Biais	EQM	Exact	Moyenne	Biais	EQM
λ_0	0,08	0,08378	0,00378	0,00046	0,08	0,08652	0,00652	0,00032
E_a	0,88	0,90365	0,02365	0,05653	0,05	0,19523	0,14523	0,04766
η	1,23	1,02521	-0,20479	0,30846	1,79	1,56142	-0,22858	0,33988
λ_0	0,08	0,08737	0,00737	0,00039	0,08	0,08866	0,00866	0,00075
E_a	0,1	0,22461	0,12461	0,04501	2,1	1,68384	-0,41616	0,25014
η	2,56	2,34348	-0,21652	0,39207	2,95	2,57029	-0,37971	1,81504
λ_0	0,08	0,09250	0,01250	0,00061	0,88	0,09260	0,01260	0,00077
E_a	1,17	1,08988	-0,08012	0,05193	1,83	1,52107	-0,30893	0,16340
η	3,35	2,88426	-0,46574	0,92119	3,47	2,91897	-0,55103	1,61020

FIG. 1 – *Taux de défaillance d'éléments posés à différentes dates pour le modèle ajusté sur les données réelles* ($\lambda_0 = 0,00024, E_a = 0,61481, \eta = 2,41268$)

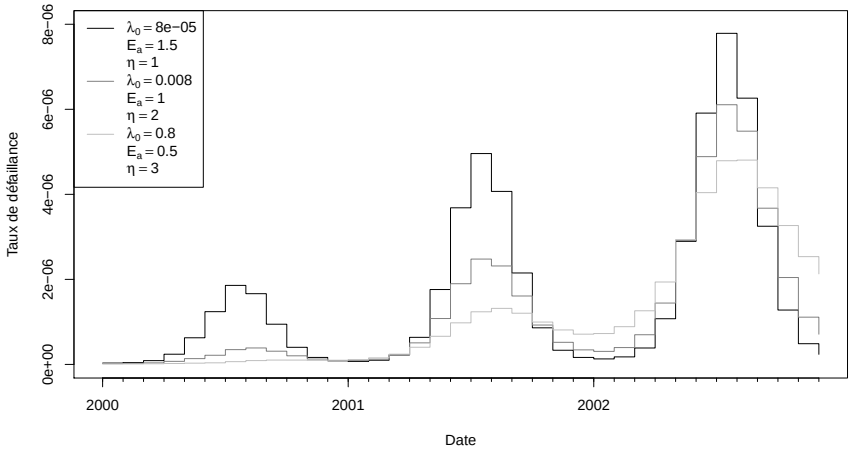


4 Remarques et Conclusion

L'estimation des paramètres E_a et η permet de quantifier l'effet de la température et celui de l'humidité :

- E_a affecte le contraste entre la fiabilité des mois chauds et celle des mois froids. Plus E_a est élevé, plus ce contraste sera important (Fig.2),
- η joue sur la tendance au vieillissement. Cette tendance croîtra d'autant plus rapidement que η est grand (Fig.2).

FIG. 2 – *Taux de défaillance pour des éléments installés au début d'une étude de trois ans, pour différentes combinaisons des paramètres (simulations)*



Nous pouvons imaginer étendre ce modèle à d'autres facteurs environnementaux suivant l'équipement à étudier. Nous sommes alors en mesure de connaître l'effet environnemental prépondérant sur le mécanisme de panne, ou encore de prédire la réaction des matériels sous différents scénarios d'environnement. Ce modèle physico-statistique offre aux industriels une meilleure visibilité sur le comportement de leurs équipements face aux agressions climatiques. Ces informations peuvent être utiles dans l'élaboration d'une politique de maintenance préventive.

Bibliographie

[1] Carer, P., Lattes, R., Charles, J.C., Puluhen, B., Zorzi, F., Espilit, T. et Pierrat, L. (2010) Using physical and statistical Cox models for reliability estimation of aged electrical components subjected to environmental stresses. *International Conference on Accelerated Life Testing, Reliability-based Analysis and Design.*
[2] Crank, J. (1979) *The Mathematics of Diffusion*, Oxford University Press. Seconde édition.
[3] Gouno, E. (2007) Optimum step-stress for temperature accelerated life testing. *Quality and Reliability Engineering International*, 23, 915–924.
[4] Kim, J. et Proschan, F. (1991) Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability*, 40, 134–139.
[5] Lall, P. (1996) Tutorial : Temperature As an Input to Microelectronic-Reliability Models. *IEEE Transactions on Reliability*, 45, 3–9.
[6] Monroe, E. et Pan, R. (2008) Knowledge-based assessment for time varying climates. *Quality and Reliability Engineering International*, 25, 111–124.
[7] Peck, D. (1986) Comprehensive Model for Humidity Testing Correlation. *24th Annual Reliability Physics Symposium*, 44–50.
[8] Tencer, M. (1994) Moisture ingress into nonhermetic enclosures and packages. A quasi-steady state model for diffusion and attenuation of ambient humidity variations. *IEEE 44th Electronic Components and Technology Conference.*

Détection de la défaillance des entreprises tunisiennes par la régression logistique semi paramétrique et les réseaux de neurones

Abdeljelil Farhat & Sami Mestiri

Unité de recherche EAS-Mahdia
Faculté des sciences économiques et de gestion de Mahdia,
Université de Monastir, Tunisie.

Résumé : L'objectif de cet article est de comparer deux techniques de classification des entreprises : la régression logistique semi paramétrique et les réseaux de neurones dans le but de prévoir le risque de crédit des banques tunisiennes. L'échantillon utilisé comporte 528 firmes tunisiennes de différents secteurs d'activités dont nous disposons des bilans et des comptes financiers des exercices 1999-2006. Nous avons démontré que l'examen soigneux du rapport fonctionnel entre les ratios et la probabilité de la détresse et l'emploi de modèles basés sur les réseaux de neurones améliore la qualité des prévisions des modèles en terme de bon classement.
Mots clés : Prévision ; Risque de crédit ; Régression logistique semi paramétrique ; Réseaux de neurones ; Courbe ROC.

Abstract : The aim of this paper is to compare the forecasting financial distress models named the the semi parametric logistic scoring and neural networks. Financial variables are used to predict financial situation of Tunisian firms. The semi parametric logistic scoring is established by the careful examination of the functional relationship between the ratios and the probability of the distress. The proposed semi parametric model has a better discriminatory power and a predictive performance.
Key words : Forecasting ; Logistic semi parametric model ; Neural networks ; Curve ROC.

Introduction et sommaire

La prévision de la détresse financière d'entreprises est une procédure très importante pour ceux qui y sont impliqués (actionnaires, gestionnaires, salariés, prêteurs, fournisseurs, clients et surtout l'État). Les modèles de prévision servent comme "système d'alerte " pour les gestionnaires d'entreprises qui peuvent entreprendre des actions de prévention contre le risque de faillite (par exemple, opération de rachat, de liquidation, de redressement, etc.). D'autre part, ces modèles peuvent être aussi utiles pour les professionnels des établissements financiers dans l'évaluation et la sélection des entreprises auxquelles ils prêtent des crédits. En partant de ces considérations et devant l'ampleur du phénomène, diverses études et recherches ont été menées dans ce sens durant ces trente dernières années. Elles visaient à mettre en évidence les principaux indicateurs permettant de prévoir à temps les

difficultés éprouvées par les entreprises. Nous pouvons citer parmi les premiers travaux, à titre d'exemple, ceux de Beaver (1966) et Altman(1968).

Depuis cette période et jusqu'à nos jours, le nombre des études sur l'évaluation des risques de faillite et la prévision de la détresse financière des entreprises ne cesse d'accroître. Il suffit de citer Bardos et Zhu (1997), Chava et Jarrow (2004) et Hillegeist (2004). La grande majorité de ces recherches s'appuie sur des outils d'analyse statistique de grandeurs comptables et de ratios financiers pour discriminer les entreprises saines des entreprises défaillantes. Ces études ont abouti à une fonction de score qui est un indicateur de synthèse censé de donner en un chiffre, le degré de défaillance possible d'une entreprise.

Ce papier s'intègre dans le cadre de comparaison deux techniques de classification des entreprises : la régression logistique semi paramétrique et les réseaux de neurones. Ainsi, nous avons réalisé une recherche exploratoire des nouvelles relations fonctionnelles entre les ratios et la probabilité de la détresse. Ces relations fonctionnelles ont été estimées à travers le modèle de régression logistique semi paramétrique. D'un autre coté, nous avons appliqué la technique des réseaux de neurones artificiels à la prévision de la détresse financière des firmes tunisiennes.

Dans une étude relative à des entreprises américaines, Press et Wilson (1978) ont utilisé des données de ratios en coupe transversale pour examiner si les coefficients de la fonction de score estimés à partir du modèle de la régression logistique sont des déterminants valides de la faillite des entreprises. Une caractéristique importante de ce modèle régression logistique est la liaison paramétriquement de la moyenne conditionnelle de la variable expliquée est liée aux variables explicatives.

En réalité, l'hypothèse que la forme fonctionnelle dans le modèle de régression est linéaire souvent n'est pas appropriée surtout lorsque le phénomène étudié est compliqué. Pour contourner cette lacune, Zhang et Lin (2003) ont proposé une modélisation flexible des effets des variables explicatives ou le prédicteur linéaire dans le modèle de régression est remplacé par des fonctions non paramétriques. Le nouveau modèle est nommé le "Modèle de régression logistique semi paramétrique". L'intérêt principal du modèle semi paramétrique est qu'il permet de distinguer les relations linéaires et les relations non linéaires au sein d'un même modèle.

Dans notre application, nous avons choisi de retenir 8 ratios significatifs liés aux différentes dimensions de l'analyse financière et qui représentent les différents critères d'appréciation de la bonne santé d'une entreprise. Sur la base de des constats tirées a partir des présentations graphiques de nuages des points des ratios en fonction des rapports de chances correspondants, il est intéressant de considérer une modification de la variable R_{21} dans le modèle de régression. Ainsi, nous considérons ce modèle de régression logistique semi paramétrique, qui s'écrit sous la forme suivante :

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 R_{7i} + \beta_3 R_{9i} + \beta_4 R_{10i} + \beta_4 R_{14i} + \beta_5 R_{15i} + \beta_6 R_{20i} + f(R_{21i}) \quad (1)$$

Avec $p_i = P(y_i = 1|R_i)$, pour $(i = 1, ..., n)$ est la probabilité a posteriori d'appartenance au groupe d'entreprises en détresse, β est un coefficient inconnu et f est une fonction de lissage inconnue.

Selon l'approche de Wand et Ngo (2004), l'estimation du modèle de régression logistique semi paramétrique (1) revient à estimer le modèle de régression logistique à effets aléatoires en supposant que le vecteur des effets aléatoires b soit normalement distribué $N(0, G_\theta)$.

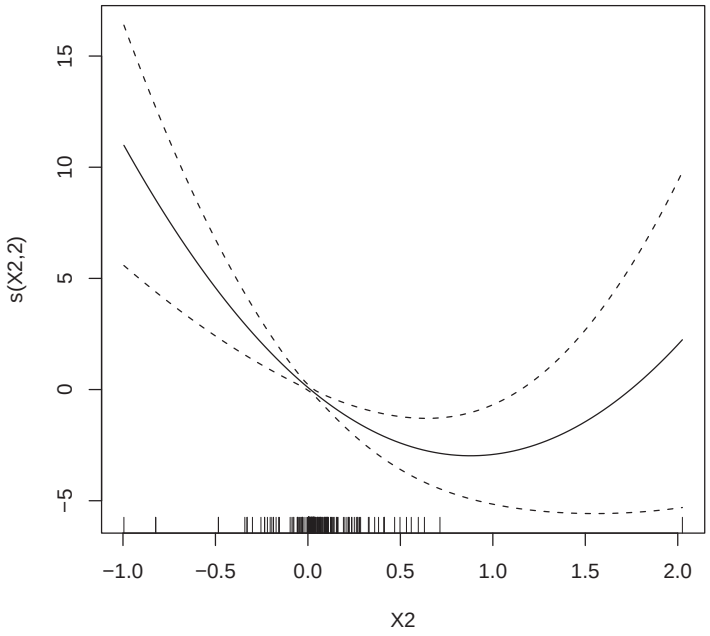


FIG. 1 – La courbe de la fonction estimée $f(R_{21})$

La figure (1) représente la courbe de la fonction $f(R_{21})$ estimée de la variable capacité d'endettement à long terme avec son intervalle de confiance à 95%. D'après ce graphe, pour un seuil inférieur à 1, la probabilité de détresse est une fonction décroissante de la capacité d'endettement à long terme et pour un seuil supérieur à 1, elle devient croissante.

	Le réseaux de neurones			La régre. logistique semi		
	$\hat{Y}=1$	$\hat{Y}=0$	Total	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Y=1$	24	14	38	6	3	9
$Y=0$	65	510	575	83	521	604
Le taux d'erreur	0.128			0.140		

TAB. 1 – Matrice de confusion des modèles estimés pour l'échantillon test

Si l'analyse par le modèle de régression logistique semi paramétrique est devenu connu comme une procédure économétrique caractérisées par deux étapes (la création d'un modèle suivie par l'estimation de ses paramètres), les réseaux de neurones appartiennent à une catégorie différente d'outils d'analyse des données. En effet, Tam et Kiang (1992), Altman (1994), ont intégré les progrès enregistrés en matière d'intelligence artificielle pour la construction des modèles de prévision de la détresse financière des firmes. Les modèles développés sont non linéaires et non paramétriques et tiennent compte des avancées effectuées en matière de reproduction artificielle des réseaux de neurones et de génétique.

Dans le contexte de notre travail, plusieurs chercheurs (Perez, 2006 ; Chih-Fong et Jhen-Wei, 2008 ;Tilmont, 1998 ; Bardos et Zhu, 1997) ont proposé l'emploi de modèles basés sur les réseaux de neurones en vue d'améliorer la prise de décision du banquier.

Étant donné que les modèles de réseaux de neurones se construisent par apprentissage à partir d'un certain nombre d'observations. Tout au long de cette application, nous avons utilisé uniquement 80% des observations pour la Formation (Apprentissage) et les 20% restant afin de tester la capacité prédictive réelle du réseau.

Le réseau a été entraîné sur tout l'ensemble d'apprentissage correspondant aux 8 ratios calculés. Pour chaque configuration à tester, le réseau de neurones essaie de déterminer l'ensemble des pondérations optimales des inputs. Les valeurs des outputs obtenus, grâce à chaque système de pondération, sont comparées au statut réel de l'entreprise : saine ou en détresse. Ce statut étant appréhendé à travers un score égal à 1 lorsque l'entreprise est saine et à 0 dans le cas contraire. Le pourcentage de classifications correctes des firmes est calculé en rapportant le nombre d'observations correctement classées au nombre total d'observations dans l'échantillon d'apprentissage.

Après avoir déterminé des fonctions de score de la détresse, il faut en évaluer leurs efficacités. Nous pouvons le faire par les tests du pouvoir discriminant et les tests du pouvoir prédictif. Ainsi, nous allons calculer le taux d'erreur de classement et tracer la courbe de ROC "Receiver Operating Curve" en calculant les indices associés tels que l'aire sous la courbe de ROC.

Nous rappelons que le taux d'erreur de classement est égal au nombre de mauvais classement rapporté à l'effectif total. D'après la table 1, Le taux d'erreur de classement

égale à 14% pour le modèle de la régression logistique semi paramétrique et 12.8% pour les réseaux des neurones c.à.d une amélioration de prédiction de 1.2%. Ce qui prouve l'importance de l'utilisation de la technique des réseaux des neurones dans le calcul de risque de la détresse.

La courbe ROC est un outil graphique qui permet d'évaluer et de comparer globalement le comportement des fonctions de scores (Pepe(2000)). Elle est indépendante des coûts de mauvaise affectation. Elle est aussi opérationnelle même dans le cas des distributions très déséquilibrées. Mieux, même si les proportions des classes ne sont pas représentatives des probabilités a priori, la courbe ROC reste valable.

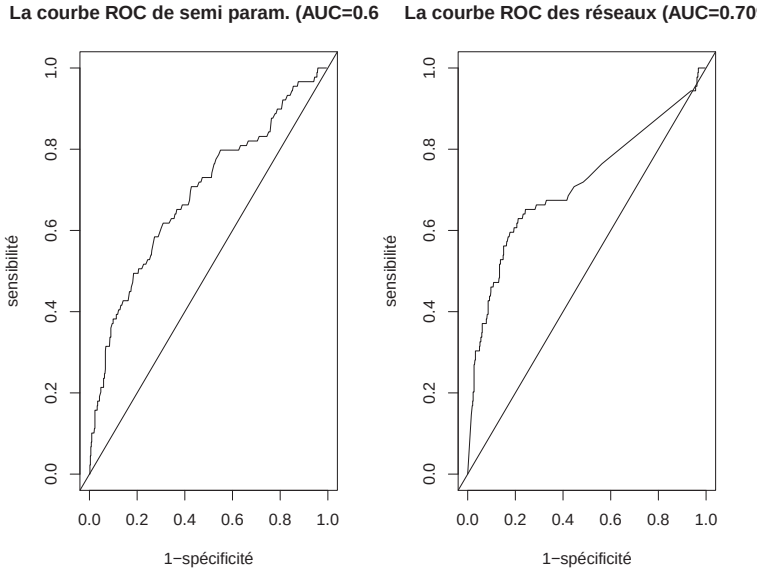


FIG. 2 – Les courbes ROC des modèles établis

D'après la figure 2, il est évident que la règle de classification basée sur les réseaux de neurones est plus performante que celle basée sur la régression logistique semi paramétrique. Ceci nous amène à conclure que la validité de la fonction de score issue du modèle de réseaux de neurones soit meilleur que celle obtenue à partir du modèle de régression logistique semi paramétrique.

L'aire sous la courbe roc (AUC) mesure la qualité de discrimination du modèle et traduit la probabilité qu'une entreprise saine ait un score supérieur au score d'une entreprise en détresse, ceux-ci étant tirés au hasard. L'AUC du modèle de régression logistique semi paramétrique est égale à 0.61 par contre l'AUC de réseaux de neurones est égale à 0.709 ; plus sont proches de un.

Nous avons visé plus particulièrement à attirer l'attention sur l'aspect non linéaire des relations entre les ratios et la probabilité de la détresse. D'autre part, montrer que les réseaux de neurones artificiels est un outil de prévision puissant en matière de détresse financière des firmes.

Bibliographie

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4) :589-609.

[2] Altman E.I, Marco G. and Varetto F. (1994), Corporate distress diagnosis : comparisons using linear discriminant analysis and neural networks : the Italian experience, *Journal of banking and finance*, vol. 18 n°3, pp. 505-529.

[3] Bardos, M. and Zhu, W. H. (1997). Comparaison de l'analyse discriminante linéaire et des réseaux de neurones. application à la détection de défaillance d'entreprises. Revue Statistique Appliquée.

[4] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4 :71-111.

[5] Breslow, N. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models". *Journal of the American Statistical Association* 88 :9 - 25.

[6] Chava, S. and Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects. *Review of Finance*, 8(4) :537-569.

[7] Tam K.Y.et Kiang M.Y. (1992), Managerial application of neural networks : the case of bank failure predictions, *Management science*, vol.38 n°7, pp.926-947.

[8] Ngo, L. and Wand, M. (2003). "Smoothing with mixed model software". *Journal of Statistical Software*, 4(1) :1-54.

[9] Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449) :308-311.

[10] Press, S. J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364) :699-705.

[11] S. Hillegeist, E. Keating, D. C. and Lundstedt, K. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9 :5-34.

[12] Zhang, D. and Lin, X. (2003). "Hypothesis testing in semiparametric additive mixed models".*Biostat*, 4(1) :57-74.

Fuzzy multivariate cumulative sum & exponentially weighted moving average control chart

†ALI ACHOURI & *HASSEN TALEB

†*To whom correspondence should be addressed, e-mail:ali.achouri@live.fr*
Institut Supérieur de Gestion, Department of Statistics. University of Tunis, LARODEC. 41. rue dc la liberté- cité Bouchoucha- le Bardo-2000-Tunisia.
*Associate Professor; Quantitative methods department; Higher Institute of Business Administration of Gafsa, University of Gafsa

Abstract

In this article, fuzzy set theory is proposed to construct multivariate cumulative sum (cusum)& exponentially weighted moving average (ewma)control chart. Real data taken from "Idéal Sanitaire" process is used. Product units are classified into 5 categories. Each sample is presented by a single fuzzy set and each category is described by a linguistic variable. The resulted control charts are compared and analysed.
keywords: Statistical process control, fuzzy set, linguistic variable, multinomial data, membership function, representative value.

Résumé

Dans cet article, la théorie floue est utilisée pour construire les cartes de controles de la somme cumulée multi variée et de la moyenne pondérée exponentielle multi variée. Des données réelles de la société "Idéale Sanitaire"sont utilisées. Les produits sont classés en 5 catégories et chaque échantillon est résumé par une variable linguistique.
Mots clés: controle statistique des processus, ensembles flous, variable linguistique, données multinomiales, fonction d'appartenance, valeur représentative.

1 Introduction

Montgomery (2003) explained the various steps of construction of cusum and ewma control charts. Charts are constructed on the basis of numerically measurable data. Futoshi et al(1996) required human subjectivity and probability to construct the membership function. Taleb and Limam (2002) compared several procedures for constructing control charts based on fuzzy theory and linguistic variables, Zadeh(1965,1975). Each quality category is associated with one fuzzy set and each sample was represented by fuzzy representative value. Alipour and Noorossana (2010) proposed an application of fuzzy theory on multinomial data to draw ewma control charts. We propose to apply fuzzy theory

to linguistic and multinomial data to construct cusum and ewma control charts based on real data to compute representative values is proposed. An overview of the approach and membership function construction will be proposed . Cusum & Ewma control chart are discussed and real data example is proposed to apply the new method. Fuzzy set is composed by objects or characters with different degrees of belonging in it.

$$\begin{array}{c} \mu_A : U = [a, b] \longrightarrow [0, 1] \\ x \longrightarrow \mu_A(x) \end{array} \tag{1}$$

To resolve subjectivity due to individual uncertainty, Direct Rating is used and expert is called to answer the question "How x is u ? ". In response, the expert tries to convince him and others about the membership degree of laying the necessary arguments to validate his choice. Fot other observations, the expert must then continue the same approach for all linguistic terms or elements of x. If there are n expert, so the same approach will be repeated n times for m linguistic terms. Subsequently, for the construction of the final membership function, a linear regression is made.

2 cusum & ewma control charts

2.1 Ewma control chart

Ewma control chart is a statistical tool of control chart used to detect small variability in production processes, it presents an advantage over the Shewhart control chart that it includes at the same time all previous data added to the current .

$$Z_i = \lambda X_i + (1 - \lambda Z_{i-1}) \tag{2}$$

λ shows the weight of the actual observation to determine the statistical value of Z_i, this value is a proportion that summarizes the importance of actual observation compared by the previous ones. and $(1 - \lambda)$ is the weight of previous observations in explaining the process. If $\lambda = 0$, the current observation is removed and if $\lambda = 1$, the current observation is taken into account and old observations are removed, this case is similar to the Shewhart control chart. For small λ,in order of 0.3 or less, Ewam is effective for detecting small variability.

$$\begin{array}{l} UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1 - (1 - \lambda)^{2_i}]} \\ CL = \mu_0 \\ UCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1 - (1 - \lambda)^{2_i}]} \end{array} \tag{3}$$

L is a parameter used in the upper and lower bounds of control to show the order of sigma that we want to affect to our process. L is the multiple of standard deviation σ.

2.2 Cusum control chart

C_i is a statistic which aims to calculate the sum of deviations between the average of the observations and the target value μ_0, it gives an idea about the quality control but it is not a tool of statistical quality control because the concept of control charts requires the existence of upper and lower bounds; which don't exist in this satistic.

$$C_i = \sum_{j=1}^{i} (\overline{X_j} - \mu_0) \tag{4}$$

where μ_0 is the target value. If there is a trend upwards or downwards, there is a signal that the process has just changed its average, the value of $\overline{x}$ would be either higher or lower than μ_0 and does not change orientation, assignable cause should be detected.

$$\begin{array}{l} {C_i}^+ = max[0; X_i - (\mu_0 + k + {C_{i-1}}^+] \\ {C_i}^- = max[0; (\mu_0 - k) - X_i + {C_{i-1}}^-] \end{array} \tag{5}$$

k is a parameter, it is called a reference value, it is equal to the average of the target value μ_0 and the new average μ_1 $K = \frac{\mu_0 + \mu_1}{2}$. ${C_i}^+$ and ${C_i}^-$ are cumulative deviations from the target value μ_0, such that $C_i^{(+/-)} \subset R^+$. If we find a negative difference, so it is essential to bring them to 0. So the logic of this method is essentially to draw deviations ${C_i}^+$ and ${C_i}^-$ around μ_0, and then see if the process is controlled or not by comparing the upper and lower bounds.
The control limits are defined by H knowing that H is a multiple of σ. If ${C_i}^+$ or ${C_i}^-$ exceeds H , so the process is out of control, we must seek the time from which the process becomes out of control. To find the time t source of deviation , we must calculate N knowing that N indicates the number of consecutive periods where C_i is strictly positive. The first period where $N > 0$ is the period from which there is an assignable cause that has triggered the process, the objective is to detect and remove it.

3 Numerical example

"Idéal Sanitaire"is a multinational company established in 1992 and available in over 44 countries in the world added to Tunisia specializing in sanitary appliance (washbasin, close coupled wc, kitchen sinks, Baths, Showers).

When a product has no defect or a minor defect in an invisible area, it is classified as a first choice, when it shows any minor defect in an invisible area which does not affect the use of the product, it is classified as second choice (standard). If there is a visible defect that does not affect the use of the product is classified as a third choice. Finally, when the use is affected and may be correct, the item is considered to be Repair(Class 4), otherwise waste (class 5).

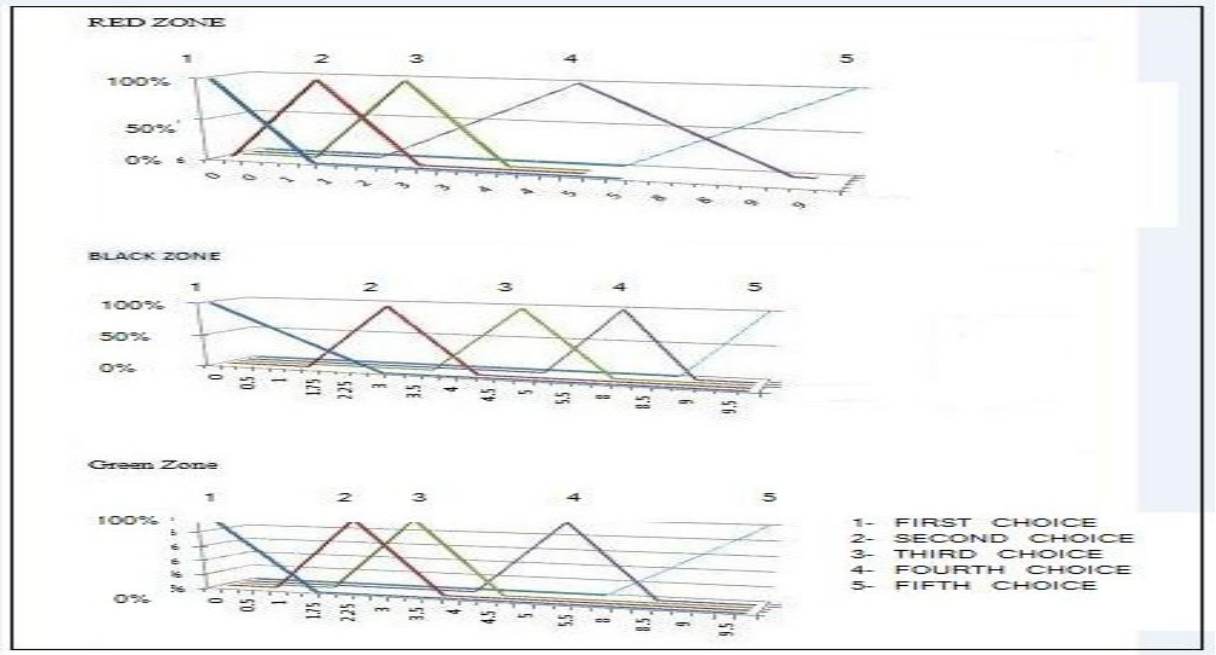


Figure 1: Fuzzy sets

Table 1: cusum & Ewma parameters

μ_0	μ_1	σ	k=$\frac{(\mu_1-\mu_0)}{2}$	$H = 5 * \sigma$
$0,0157$	$0,182$	$0,121$	$0,0834$	$0,609$
n	λ	μ_0	L	σ
40	$0,3$	$0,182$	2.7	0.121

3.1 Determination of representative values

To successfully draw fuzzy multivariate ewma & cusum control charts we have to construct the representative values, where we should firstly collect the ratings of experts to judge the category where belongs the product and transform them to membership functions (membership function 1^{st} choice, 2^{nd} choice, 3^{rd} choice, Rework and Gar), figure 1. Secondly, define a general membership function by adding different fuzzy sets (general membership function 1^{st} choice, 2^{nd} choice. ... Gar), by fuzzy addition of all zones. Then, use general membership functions to construct fuzzy sets of each sample by multiplying each category with it's proportion and make the fuzzy addition of different classes to give only one fuzzy set per sample. Finally, representative values are fuzzy modes, Table 2.

Fuzzy multivariate cusum control chart

Representative values are used to define the specific parameters of cusum control chart. C_i chart has a trend, figure 2 (trend C_i), which give an idea that the process is out of control, however, all observations lie between the limits of control chart, figure 2 (cusum control chart). If the trend continues, the process will soon be out of control.

Fuzzy multivariate ewma control chart

Z_i decreases and then becomes constant, figure 2 (plot of Z_i). All observations lie between the upper and lower bounds then the process is in control, figure 2(ewma control chart).

Table 2: Fuzzy modes per sample

sample	1	2	3	4	5	6	7	8	9	10
FM	0.22	0.15	0.24	0.17	0.47	0.04	0.29	0.27	0.06	0.10
sample	11	12	13	14	15	16	17	18	19	20
FM	0.15	0.14	0.24	0.00	0.00	0.35	0.11	0.26	0.31	0.14
sample	21	22	23	24	25	26	27	28	29	30
FM	0.14	0.25	0.15	0.26	0.29	0.26	0.28	0.21	0.35	0.31
sample	31	32	33	34	35	36	37	38	39	40
FM	0.23	0.29	0.00	0.04	0.28	0.00	0.00	0.14	0.00	0.00

4 conclusion

Fuzzy multivariate cusum and ewma control charts use current and previous observations to draw quality charts based on linguistic variables and fuzzy sets with triangular shapes and single fuzzy core .
Fuzzy sets are multiplied by proportions of quality that characterizes them in the sample and all added after that in order to get a single fuzzy set which represent all products in this sample.
Data of "idéale Sanitaire" for the 40 samples are between upper and lower limits which implies that the products are under control. However the cumulative sum shows a trend which provides that the sample will be soon out of control. Future research should study accurately the exact ways of membership function's identification to reduce subjectivity and improve the use of mathematical tools.

Bibliographie

[1] Alipour.H.,Noorossana.R(2010); "Fuzzy multivariate exponentially weighted moving average control chart"; 48:10011007 DOI 10.1007/s00170-009-2365-4.
[3] Futoshi,T.,Akihiro,K.,Hiroshi,O.1998 "Identification of membership functions based on fuzzy observation data",311-318
[4] Montgomry,D. 2003 ,"Introduction To Statistical Quality Control",third edition.
[5]Taleb.H.,Limam,MMT,(2002);"On fuzzy and probabilistic control charts"; 40;12. 2849-2863.
[6] Zadeh.L.A.(1975);"The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I";Information Sciences 8,199-249.
[7] Zadeh.L.A.(1975); "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-II";Information Sciences 8,301-357.
[8]Zadeh.L.A.(1965);"Fuzzy sets";Information and control 8, 338-353

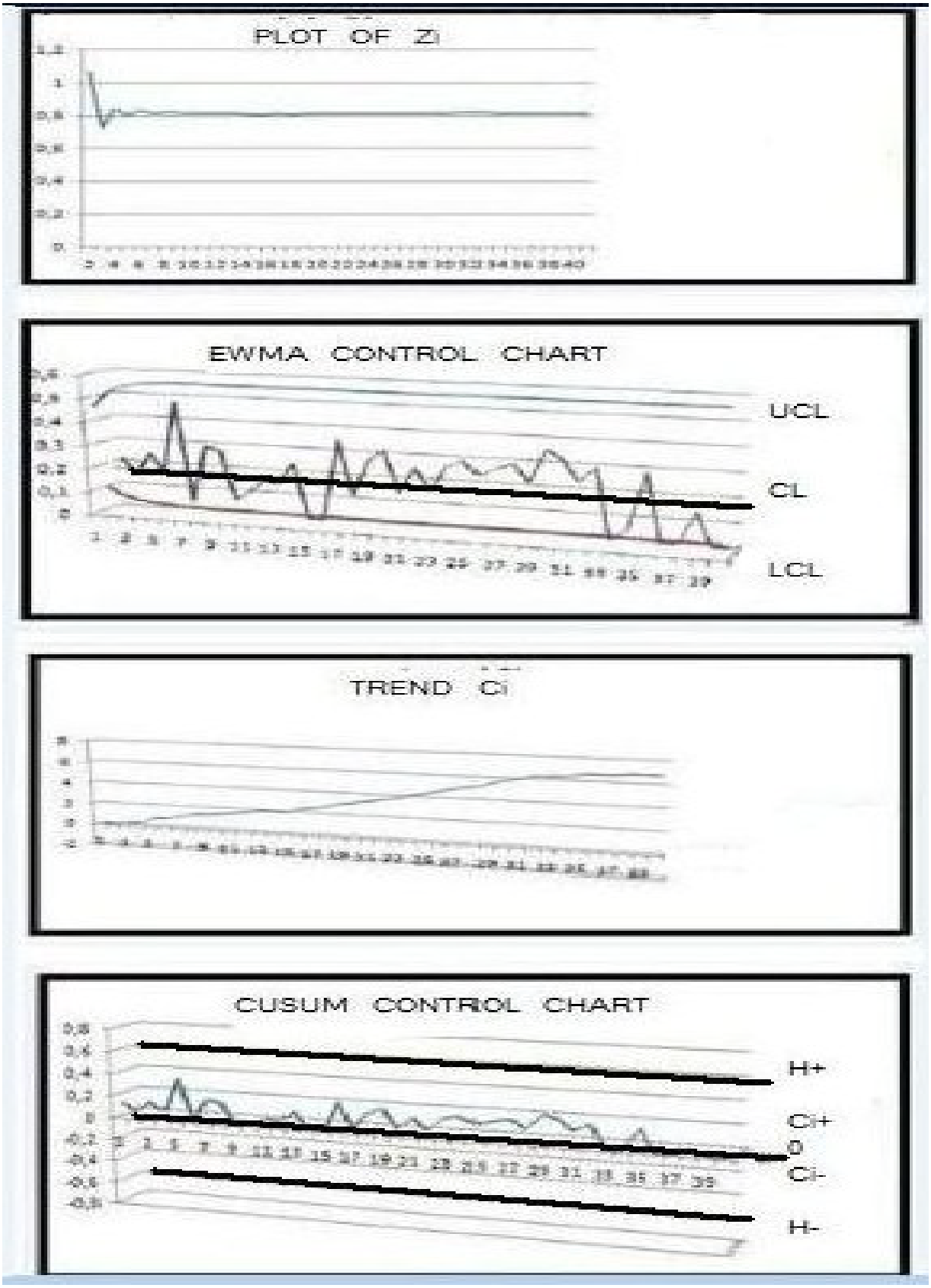


Figure 2: Fuzzy Mewma & Mcusum

Multiscale process monitoring in Nanomanufacturing

Sihem Ben zakour*& Hassen Taleb+

**To whom Correspondence should be addressed, e-mail: benz_sihem@yahoo.fr, Higher institute of management, Department of Statistics. University of Tunis, LARODEC + Associate Professor, Quantitative methods department Higher Institute of Business Administration of Gafsa*

Abstract

The detection of the end of polishing during the chemical mechanical planarization (CMP) process is a critical task in semiconductor manufacturing. The disadvantages of offline approach has incited the researchers to discover an efficient substitute. In this paper, an alternative approach named online method has been presented in which a sequential probability ratio test (SPRT) was developed and applied to the wavelet decomposed Acoustic emission data collected during the progression of the CMP process. This test is shown to be efficient in controlling complex processes and appropriated for real-time application by developing a moving block strategy.
Keywords: Sequential probability ratio test, End point detection, Chemical mechanical planarization, Acoustic emission, Wavelet-Based Multiresolution Analysis.

Résumé

La détection de la fin du polissage au cours de la planarisation mécano-chimique (CMP) est une tâche critique dans la fabrication de semi-conducteurs. Les inconvénients de l'ancienne approche a incité les chercheurs à découvrir un substitut efficace. Dans cet article, une approche alternative a été présentée sous le nom de méthode en ligne dans laquelle un test de ratio de probabilité séquentiel (SPRT) a été développé et appliqué à des données acoustiques recueillies au cours de la progression du processus de CMP. Ce test a montré son efficacité dans le contrôle des processus complexes et dans l'application en temps réel par lélaboration d'une stratégie de bloc mobile.
Mots-clés: Test de ratio de probabilité séquentiel, Détection de point de fin de polissage, Planarisation mécano-chimique, Émission acoustique, Analyse d'ondelettes basée sur la multirésolution.

1 Introduction

The CMP process combines a mechanical and chemical propreties allowing to ensure the wafer flatness. The reduction of wafer size, the introduction of new material like the copper and low-k dielectrics and the superposition of differents metals create an enormous challenge in terms of accomplishment of wafer production and to check its quality. Hence, it seems necessary, even crucial to determine the right moment of stopping this process using AE sensor. In this paper, Sequential probability ratio test (SPRT) for variance and coefficient of variation applied on the wavelet decomposed AE data will be presented, with comparison between these two tests will be conducted to identify the most powerful test. The organization of this paper is as follow: In section II, detailed description about the online monitoring using SPRT will be proposed. The experimental setup and results will be given in section III. In section IV, presents the concluding remarks.

2 Online monitoring using SPRT

Abraham Wald(1973) defined a parametric statistical test named sequential test where the number of data collected is not predetermined and treated as a random variable. This method is developed especially for monitoring the quality of a specific product. The sequential method of testing a hypothesis H has three decisions at any stage of experiment:
(1)To accept the hypothesis H_0
(2)To reject the hypothesis H_0
(3)To continue the experiment by making an additional observation
Neyman and Pearson (1933) have presented a new method to construct the most powerful test for simple test. Supposing X has p.d.f f (X,θ), the hypothesis test $H_0 : \theta = \theta_0$ or $H_1 : \theta = \theta_1$. The base of test is the Neyman-Pearson (N-P) Lemma , which states that, for a fixed sample size of (n), the optimal design (the most powerful test) for simple hypothesis (H_0 angianst H_1) can be obtained from the likelihood ratio (λ_n) is given by:

$$Accept H_0 \quad if \ \lambda_n \prec k \tag{1}$$

$$Reject H_0 \quad if \ \lambda_n \geq k \tag{2}$$

$$\lambda_n = \prod_{i=1}^{n} \frac{f(X_i \,|\theta_1)}{f(X_i \,|\theta_0)} \tag{3}$$

k is the decision limit associated with level of significance α (the critical region size), and (i) denotes the observation index. The distribution of X_i must be Gaussian for implementing SPRT because this test is not applied to the original data but to the wavelet details (which is the X_i). Applying SPRT on the variance and CV of wavelet details, the upper control limits are needed to detect the end point. The design parameters of the

Table 1: SPRT for variance and coefficient of variation

	For variance	For coefficient of variation
Reject H_0(Accept H_1: End point reached)	$\sigma_n^2 \geq A$	$\frac{\sigma_n}{\mu_n} \geq \frac{\sqrt{A}}{\mu}$
Keep on sampling	$\sigma_n^2 \prec A$	$\frac{\sigma_n}{\mu_n} \prec \frac{\sqrt{A}}{\mu}$

SPRT chart are σ_0,σ_1,α,and β. And, the equations A, σ_0 and σ_1 are defined as follow:

$$A = \frac{2\log(\frac{1-\beta}{\alpha}) + n\log(\frac{\sigma_1^2}{\sigma_0^2})}{(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})n} \tag{4}$$

$$\sigma_1 = \overline{S} + 2\overline{S}\frac{(\sqrt{1-c_4^2})}{c_4} \tag{5}$$

$$\sigma_0 = \overline{S} + 1.5\overline{S}\frac{(\sqrt{1-c_4^2})}{c_4} \tag{6}$$

With c_4 is a constant and depends only on the value of n

$$c_4 = {}^{4(n-1)}\!/_{(4n-3)} \tag{7}$$

$\bar{S}$ is the mean value of the standard deviation of wavelet details. And the choice of the values 2 and 1.5 are made according to the simulated data.

3 Online moving block monitoring strategy in SPRT

The online methodology with multiscale analysis contains two stages: Firstly, the application of wavelet based multiresolution analysis in online strategy, and using the data analyzed previously to control the progression of CMP process keeping it in the real time implementation. The implementation of sensors allows to measure and to convert the output of data and the environmental factors into signal subsequently to transmit the

obtained signal to data acquisition system. The PC based DAQ system is charged of the acquisition of electrical signals and digitized it. To make interface between data acquisition system (DAS) and MATLAB, the use of an appropriate scientific engineering software will be unavoidable allowing the transfer of the data digitized by DAS at a particular dyadic length, to a (.m) file in MATLAB. The obtained Matlab file is considered like the first block of input. The first dyadic data block is wavelet decomposed into coefficients at the appropriate level based on data type after that denoised using thresholding methods and reconstructed into the time domain wavelet details. Standard deviation of the wavelet details in the first data block is calculated and affected to $\overline{s}$. The value of $\overline{s}$ will be used to calculate σ_0 and also σ_1. The mean also should be accounted in order to determine the coefficient of variation. Hence, the two SPRT limits are specified. At each block containing the new collected data to be monitored, the variance and coefficient of variation of this wavelet detail are defined and plotted them against the SPRT limits. The standard deviation of each point in new block is equal to the standard deviation of all the previous wavelet details until the current point. At each new formation of data block, one must always up-date the standard deviation and coefficient of variation of the wavelet details. The value of $\overline{s}$ will be equal to the average of the standard deviations of the current and all of the past blocks. And the values of σ_0 and σ_1 will be changed according to the new value of $\overline{s}$. Hence, the upper SPRT limit of CV and variance are also up-dated. As, the upper control limit for every block is drawn from the data itself, allow us to have robust limits against all the fluctuations in the details. When the desired event occurs (correspond on the acceptance of H1), that is mean the increased value of the variance of the wavelet details exceed the upper control limit of the SPRT chart. And, if it exceed the upper limit of the control chart, will indicate the beginning of the end of planarization. For more details about the online methodology, read Das et al.(2005), Ganesan et al.(2008).

4 Description of experimental setup and results

The test bad should be equiped with an AE sensor in order to collect the data during CMP. The raw signal is nonstationary since the mean value change during the progression of CMP process. The planarization of the wafer is done under a specific combinations of rotational speed (rpm) and downward pressure (psi). In this paper, the studied data are simulated. A sample of these data is plotted in figure 1 shows the presence of noise which should be eliminated to have a robust statistical tool used in monitoring the CMP process, the amplitude of the AE signal tends to decrease during the progression of polishing procedure allowing to better situated the status of wafer(underpolished, desired level of polishing, overpolished) and the presence of white noise and autocorrelation in AE signal, the use of wavelet analysis proving its necessity.

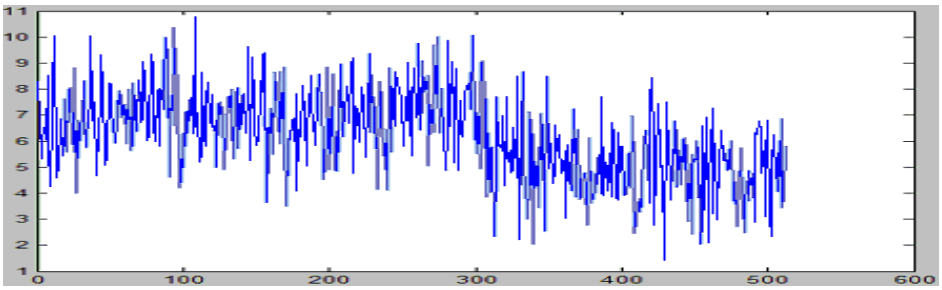


Figure 1: The representation of the simulated raw signal

To monitor the polished wafer, an online approach has been implemented. This strategy employs specific segments of generated data collected during the progression of the CMP process and called moving block. When the CMP progresses and the collected data length equals the selected window width, the analysis begins. The denoising procedure is done for the data in the first block. The wavelet chosen is daubechies fourth-order wavelet basis functions (db4) since it is more widespread and practicable in the discrete wavelet analysis. The highest energy value, which corresponds on the depth of decomposition, is equal to 8. Hence, the dyadic block width should be equal to 8 (contains 256) as the minimum length value. And the multiresolution analysis is restricted to eight levels of decomposition. At each scale of decomposition, the high frequency filtered out showed by the detail coefficient. The original signal will be approximated through the use of the set of the basis functions (wavelets) of the first scale in order to give the approximation coefficient a1 and the coeffcieint detail d1 represents the difference between the first approximation signal a1 and the original signal. The same work is done until the last level of decomposition 8. It should be noted that, the removing of the high frequency components (details), leads to significant change in the characteristics of signal and removing the low frequency (approximations), will be have no significant change in the signal. Hence, the approximation coefficients do not contain any significant informations. Therefore, they will be not used in the control procedure. The length of block n=256 (2^8) is chosen. The detail coefficients must be soft thresholded using threshold rule of Donoho and al. (1995) to only extract the significant coefficients. The reconstruction of the thresholded wavelet coefficients is unavoidable to locate the exact positioning of the happening of the end point. The values of α and β are equal to 0.1 and 0.1 respectively. The application of SPRT for variance and for coefficient of variation is done only to the details. The standard Deviation is a measure of how closely a series of numbers tracks its expected value and the coefficient of variation measures the spread of a set of data as a proportion of its mean. The acceptance of H1 is verified on the 231 point for variance SPRT and 202 collected point for the coefficient of variation. The SPRT for the first block are plotted. It is useless to form a new block.

5 Conclusion

Due to the occurence of various events in a process at different time and frequency localizations, the most industrial processes are represented by data situated at multiple scales. Hence, the application of wavelet analysis in the system identification has become ineluctable. And the best strategy for monitoring the quality of the product at nanoscale is the online strategy based on Acoustic emission sensors. This strategy is wirespread due to the high cost of ownership needed by offline method. In this paper, the acoustic signal has been analyzed using wavelet based multiresolution after that an application of SPRT for variance and CV to the reconstructed details coefficients will be employed to detect the end point. These tests show their performance in detecting the end point. But, CV SPRT allows to detect the starting of the end point of polishing earlier compared to variance SPRT. And this difference in the time of detecting the end point is not very significant. The suggested research in the futur, are ARL performance of the different measurements of EPD, examination of the on-line performances using wavelet-based multiscale to monitor differents types of defects in CMP processes.

Bibliographie

[1] Das T., Ganesan R., Sikder A., and Kumar A. (2005) *Online End Point Detection in CMP Using SPRT of Wavelet Decomposed Sensor Data*, IEEE transactions on semiconductor manufacturing, VOL. 18, NO. 3.

[2]Donoho,D.L. and Johnstone, I.M. (1995) *Adapting to unknownsmoothness via wavelet shrinkage*, Journal of the American Statistical Association, 90(432), 12001224.

[3]Ganesan R., A. N. V. Rao, and Das T. (2008) *A Multiscale Bayesian SPRT Approach for Online Process Monitoring*, IEEE transactions on semiconductor manufacturing, VOL. 21, NO. 3.

[4] Neyman J., Pearson E. (1933). *On the Problem of the Most Efficient Tests of Statistical Hypotheses*. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231: 289337.

[5]Wald A. (1973) *Sequential analysis*, Dover, New York.

A DATA DEPTH BASED EWMA CONTROL CHART

Amor Messaoud[1], Giovanni C. Porzio[2], Hela Abidi[3], Mohamed Limam[3]

[(1)]Department of Quantitative Methods, Faculty of Juridical, Economic Sciences and Management, University of Jendouba, Tunisia
[(2)]Department of Economics, University of Cassino, Italy
[(3)]Department of Quantitative Methods, Higher Institute of Management of Tunis, University of Tunis, Tunisia

Abstract

Nonparametric control charts are attractive in many industrial applications because they do not require knowledge about the shape of the underlying distribution. A new nonparametric control chart for multivariate processes is proposed. It is based on the concept of data depth approach. Its design is discussed and it is applied to a data set from a drilling process.

Résumé

Les cartes de contrôle non paramétriques sont attrayantes dans de nombreuses applications industrielles. En effet, elles ne nécessitent pas une connaissance de la forme de la distribution des données. Dans ce travail, une nouvelle carte de contrôle multivariée et non paramétrique est proposée. Des recommandations pour sa conception et une illustration sont données. Enfin, elle a été utilisée pour piloter un processus de forage.

Key words: Statistical process control; Control charts; Data depth
Mots-clés: Maîtrise statistique des procédés ; Cartes de contrôles ; Profondeur statistique

1. INTRODUCTION

A basic assumption commonly underlying the development of control charts is that the process is well described by a normal distribution. Consequently, the statistical properties of these charts are fulfilled only if the normality assumption is satisfied. However, in practice this assumption rarely holds, and this motivates the need for nonparametric control charts. Within the literature, we found that many nonparametric charts are based on the so called data depth approach (that is, on the value of a function that measures how central a point is with respect to (w.r.t) a given in-control distribution). Although very attractive, their use has been somewhat limited by the computational efforts required to implement them in practice.

We note, though, that very recently Porzio and Ragozini (2011) proposed the convex hull probability depth, a new depth function whose empirical counterpart seems to be computationally affordable even in high dimensions. Based on that depth, with this paper we aim at proposing a new nonparametric exponentially weighted moving average (EWMA) control chart.

The paper is organized as follows. In Section 2, the proposed control chart is introduced. Its design and an illustrative example are given in Sections 3 and 4, respectively. In Section 5, an industrial application is given, while Section 6 offers some conclusions.

2. THE PROPOSED CONTROL CHART

After Liu (1995), some data depth based control charts have been introduced in the literature. Amongst these latter, Porzio and Ragozini (2007) proposed a chart based on a modified version of the convex hull peeling depth, and on a nonparametric test based on the empirical center-outward quantiles. Their work is motivated by the need of nonparametric charts that can be effectively used when the process is characterized by many variables under control. They simply introduced a Shewhart-type control chart, effective to detect large process shift. Based on their idea, we aim at improving their work by proposing a data depth based EWMA (dEWMA) chart. As known, EWMA charts are more effective to detect small process shift.

2.1 DATA DEPTH AND CENTER-OUTWARD QUANTILES

Data depth is a function that measures how deep or central a given point $X \in R^k$ is w.r.t a multivariate probability distribution F or w.r.t. a given data cloud $S = \{Y_1,...,Y_m\}$. The deepest points lie at the core of the distribution, while points with lower depths are located in the distribution tails (a general discussion on data depth is available in Liu et al. (1999), and Zuo and Serfling (2000)).

With respect to a production process, let $X \in R^k$ be the vector of quality measures to be monitored, F^0 a given in-control distribution for X, and $D(./F^0)$ a depth function defined on F^0. The depth function contours of the in-control distribution are the sets:

$$C(d) = \{X \in R^k : D(X/F^0) = d\}. \qquad (1)$$

The region enclosed by contour $C(d)$ is denoted by $R(d)$. If $R(d)$ has a probability content equal to p under F^0, and F^0 is absolutely continuous and its density function is nonzero everywhere, then depth contours are coincident with p-th center-outward quantile Q_p of F^0:

$$Q_p = \{X \in R^k : D(X/F^0) = d_p\},$$

where d_p is such that $P(X \in R(d_p)) = p.$ Center-outward quantiles define a sequence of nested convex regions of increasing depth. In the special case of F^0 belonging to the class of the elliptically symmetric distributions, Q_p's are surfaces of ellipsoids.

In the multivariate statistical process control setting, the deepest points will correspond to items of higher quality, assuming that the center of the in-control distribution is the quality target to be achieved. Therefore, w.r.t. the process, the outer-inward sequence of these center-outward quantiles defines a sequence of increasing quality levels.

Let us now assume that the depth measure of a point X w.r.t. F^0 is equal to d, $D(X/F^0) = d$. That is, X belongs to the $C(d)$ contour given by equation (1). In analogy with Porzio and Ragozini (2007), we suggest using the probability content of the corresponding region $R(d)$ as the statistic to be used to define our EWMA control chart. Let us denote by $p(X, F^0)$ this probability content. It will be estimated as illustrated in the following example.

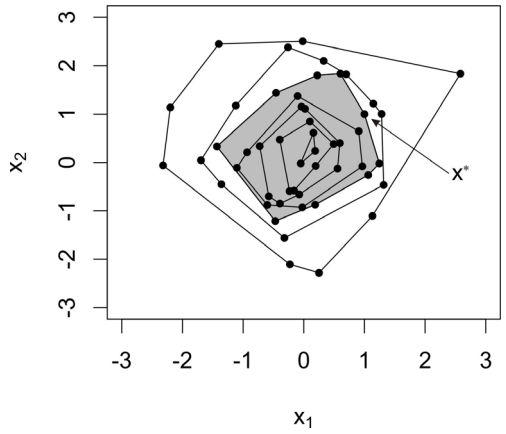


Figure 1. Illustration of the computation of $p(X^*/F^0)$.

Process data assuming F^0 is a bivariate standard normal distribution with mean vector $\mu = (0,0)'$ and with independent components have been generated. Considering the convex hull peeling sequence, the aim is to measure the probability content associated to the point $X^* = (1,1)'$ w.r.t. F^0. Let $S = \{X_1,\ldots,X_{50}\}$ be the 50 generated data points simulated from F^0, which are depicted in Figure 1. Suppose μ is the quality target to be achieved. The convex hull peeling of the sample

leads to a sequence of nested convex hulls of increasing quality levels, $CH_j(S), j = 1,\ldots,\tilde{J}$, where the index j refers to the layers, and $\tilde{J}$ is the total number of layers ($\tilde{J}$ is equal to 7 in the example). The area in gray is the area included by the convex hull to which X^* belongs in the peeling sequence, $CH_{j^*}(S)$, with j^* equal to 3 in the example. The probability content $p(X, F^0)$ is estimated as the number of vertices of $CH_j(S), j = j^*+1,\ldots,\tilde{J}$ divided by the sample size m. In the example, the observed value of $p(X^* / F^0)$ is equal to 32/50.

We note that $CH_{j^*}(S)$ is a random set because each simulated sample of size m from F^0 defines a different area, and hence corresponding probability content, $p(X^* / F^0)$, is a random number. This is why Porzio and Ragozini (2011) defined the convex hull probability depth of a point X w.r.t. F^0, as a function of the expected value of the probability content under F^0 of the convex hull to which X belongs in the peeling sequence.

2.2 THE DEWMA CONTROL CHART

Let $X_t \in R^k$ denote the vector of quality characteristic measures to be monitored at the t-th time point. It is assumed that successive X_t's are independent and identically distributed random vectors. The main idea of the dEWMA control chart is to represent each X_t by its corresponding $p(X_t / F^0)$. The control statistic of the proposed chart is thus obtained as:

$$T_t = (1-\lambda).T_{t-1} + \lambda.p(X_t / F^0),$$

$t = 1,2,...,$ where $0 < \lambda \leq 1$ is a smoothing parameter and T_0 is a starting value. Being an EWMA chart, T_0 is usually set equal to 0, assuming that the center of the in-control distribution is the quality target to be achieved. The process is considered in-control as long as $T_t < h,$ where $h > 0$ is an upper control limit. Note that the dEWMA control chart includes as a special case (when $\lambda = 1$) the Shewhart-type control chart proposed by Porzio and Ragozini (2007).

The upper-sided EWMA control chart must be considered, as the statistic $p(X_t / F^0)$ is of the type "the lower the better". That is, a decrease in $p(X_t / F^0)$ signals a process improvement: the quality of the produced item X_t will be close to the target if $p(X_t / F^0)$ is close to zero (i.e., X_t belongs to one of the deepest center-outward quantiles of F^0). Vice versa, a deterioration in the quality level is observed if $p(X_t / F^0)$ is close to one, that is if the point X_t will be in the distribution tails.

3. DESIGN OF THE DEWMA CONTROL CHART

The in-control distribution F^0 may be well defined and known. However, this is not the case in practice, and reference samples (RS) of limited amount of observations are used to describe the unknown in-control distribution. Let m denote the RS size. The design of the dEWMA control chart consists on the choice of the parameters m, λ and h according to a performance criterion of the chart. The performance of control charts is usually evaluated by the average run length (ARL), that is to say by the average number of observations that are needed to exceed the control limit for the first time. The ARL should be large when the process is statistically in-control, in-control ARL, and small when a shift has occurred, out-of-control ARL.

Crowder (1987) offered an integral equation to compute the asymptotic in-control ARL. Let $L(u)$ be the ARL of the dEWMA control chart given that $T_0 = u$. It can be shown that $L(u)$ is given by:

$$L(u) = 1 + \frac{1}{\lambda}\int_0^h L(y)\, dF\left(\frac{y-(1-\lambda)u}{\lambda}\right), \qquad (2)$$

where $F(.)$ is the distribution function of the sample statistic. According to Porzio and Ragozini

(2007), if $D(X/F^0)$ has a continuous distribution, the $p(X/F^0)$ values are uniformly distributed on [0,1] as long as the process is in-control. Consequently, the solutions to the integral equation (2) can be obtained by replacing the equation with a system of linear equations using the collocation method and solving the system of equations (Calzada and Scariano (2003)). Given that, the asymptotic in-control ARL values of the dEWMA control chart with different smoothing parameters λ and control limits h can be computed (Table 1 presents some ARL for typical values of λ and h).

Table 1. Asymptotic in-control ARL values of the dEWMA control chart with different smoothing parameters λ and control limits h.

h	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
0.50	63.402	-	-	-
0.55	119.762	-	-	-
0.60	438.528	91.537	-	-
0.65	+	311.908	52.135	-
0.70	+	+	134.655	-
0.75	+	+	591.100	109.609
0.80	+	+	+	442.479

"-" In-control ARL less than 50
"+" In-control ARL greater than 1000

In practice, we recommend the following steps in designing the proposed dEWMA control chart:

1. Choose the RS size m.
2. Choose the desired ARL (false alarm rate). The control limit h is chosen so that the control chart has this fixed in-control ARL.
3. Choose the smoothing parameter λ. Typical values are in the range $0.1 \le \lambda \le 0.3$. It is known that EWMA control charts with small (large) values of λ are more adequate to detect changes of small (large) size.
4. Compute the control limits of the dEWMA control chart.

4. AN ILLUSTRATIVE EXAMPLE

To introduce the implementation of the dEWMA control chart we offer an illustrative example. Table 2 shows 20 observations simulated from a bivariate normal distribution with mean vector $\mu = (0,0)'$ and independent components. The dEWMA control chart with $\lambda = 0.3$ and $h = 0.8$ is considered. Its asymptotic in-control ARL is equal to 442. A small RS of size 10 is considered in order to facilitate this illustration. Figure 2 shows the obtained dEWMA control chart.

Table 2. Illustrative example of the dEWMA control chart.

	Observations X_t (RS)			Observations X_t			
t	$x_{t,1}$	$x_{t,2}$	t	$x_{t,1}$	$x_{t,2}$	$p(X_t / RS)$	T_t
1	0.13	-0.09	11	0.61	-0.37	0.50	0.150
2	1.67	0.73	12	-0.29	-0.92	0.50	0.255
3	1.00	-1.28	13	-0.58	0.06	0.10	0.208
4	-2.40	-0.68	14	0.05	-0.75	0.50	0.296
5	-0.04	0.89	14	-0.14	1.48	1.0	0.507
6	-0.02	-1.30	16	-0.21	-0.26	0.1	0.385
7	-0.67	0.18	17	-0.14	-2.54	1.0	0.569
8	0.83	-0.55	18	0.58	-0.04	0.50	0.549
9	-0.64	1.01	19	-0.23	0.72	0.50	0.534
10	-0.67	-0.83	20	1.58	-0.39	0.50	0.674

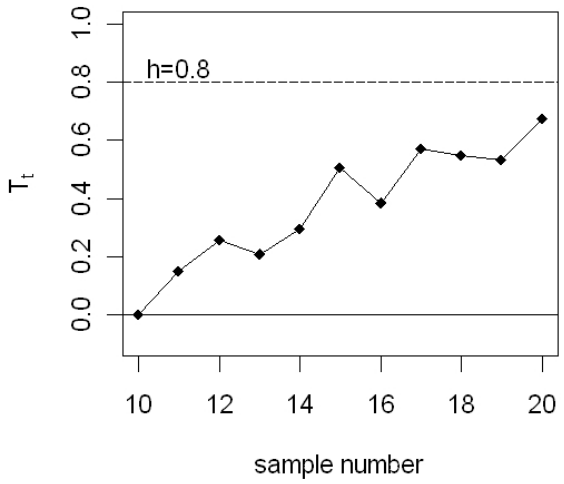


Figure 2. The dEWMA control chart (illustrative example data).

5. AN INDUSTRIAL APPLICATION

Deep hole drilling methods are used to produce holes with a high length-to-diameter ratio, good surface finish and straightness. The process is subject to chatter vibration. Chatter leads to excessive wear of the cutting edges of the tool. This reduces the machine tool life and therefore the reliability and safety of the machining operation. It may also damage the boring walls. The defect of form and surface quality constitute a significant impairment of the workpiece. Process reliability is of primary importance and hence disturbances should be avoided as the deep hole drilling process is often used during the last production phases of expensive workpieces. Therefore, it is necessary that a process monitoring system be devised to detect dynamic disturbances. In this section, the dEWMA control chart is used to detect the start of the transition from stable drilling to chatter vibration.

The data consists of the amplitudes of frequencies 234, 703 and 1183 Hz. They are the eigenfrequencies of the boring bar. They dominate the process when chatter vibration is observed. Messaoud (2006) showed that their amplitudes might be autocorrelated during stable drilling. He used a time-varying AR(1) model to remove the autocorrelation. He applied control charts to the obtained residuals to detect the transition from stable drilling to chatter vibration.

The data are obtained in an experiment where the transition from stable operation to chatter vibration starts after depth 250 mm. Chatter vibration is dominated by frequency 703 Hz and its effect is apparent on the bore hole wall after depth 300 mm. An dEWMA control chart with $\lambda = 0.3$ and $h = 0.804$ is considered. It has an in-control ARL equal to 511. This choice should not produce many false alarm signals during stable drilling. Note that at each t-th time point the RS is composed of the $m = 100$ most recent residuals of the three frequencies. Table 3 shows the out-of-control signals of the dEWMA control chart.

Table 3 shows the out-of-control signals of the dEWMA control chart. Many out-of-control signals are observed at depth 32-50 mm. They are explained by a known change in the dynamics of the process caused by the fact that the guiding pads of the drilling tool leave the starting bush. The dEWMA control chart quickly detects a change in the boundary conditions of the process at depth 225-250. This change precedes the start of the transition from stable drilling to chatter vibration that occurs after depth 250 mm. The dEWMA control chart produces 22 out-of-control signals during this transition, 250-300 mm. Chatter vibration may be avoided if corrective actions are taken after these signals, 225-300 mm. The proposed control chart shows encouraging results even though some out-of-control signals are not related to known changes in the dynamics of the process.

Table 3. Out-of-control signals of the dEWMA control chart with $\lambda = 0.3$.

Hole depth (mm)	Number of signals	Hole depth (mm)	Number of signals
≤ 32	0	175 - 200	1
32 - 50	43	200 - 225	0
50 - 75	3	225 - 250	13
75 - 100	9	250 - 275	14
100 - 125	1	275 - 300	8
125 - 150	2	300 - 325	38
150 - 175			

6. CONCLUSION

This paper introduces a new data depth based EWMA control chart. Its design is discussed and an illustrative example is given. It is then applied to monitor a drilling process. The results showed that it quickly detects the start of the transition from stable drilling to chatter vibration and that many out-of-control signals are related to known changes in the dynamics of the process.

ACKNOWLEDGMENT

The authors would like to thank the referee for making many helpful comments and suggestions on earlier versions of this article and Prof. Claus Weihs for providing the data of the drilling process.

REFERENCES

[1] Calzada, M.E. and Scariano, S.M. (2003) Reconciling the integral equation and markov chain approaches for computing EWMA average run lengths, *Communications in Statistics-Simulation and Computation*, 32, 591-604.

[2] Crowder, S.V. (1987) A Simple method for studing run-length distributions of exponentially weighted moving average charts, *Technometrics*, 29, 401-407.

[3] Liu, R.Y. (1995) Control charts for multivariate processes, *Journal of the American Statistical Association*, 90, 1380-1387.

[4] Liu, R.Y., Parelius, J.M. and Singh, K. (1999) Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion), *The Annals of Statistics*, 27, 783-858.

[5] Messaoud, A. (2006) Monitoring strategies for chatter detection in a drilling process, Ph.D Dissertation, Department of Statistics, Dortmund University of Technology.

[6] Porzio, G.C. and Ragozini, G. (2007) Multivariate control charts from a data mining perspective, In: T.W. Liao and E. Triantaphyllou (Eds.), Recent Advances in Data Mining of Entreprise Data, World Scientific, Singapore, 413-462.

[7] Porzio, G.C. and Ragozini, G. (2011) Convex hull probability depth, submitted.

[8] Zuo, Y. and Serfling, R. (2000) General notions of statistical depth function, *Annals of Statistics*, 28, 461-482.

Modèles de Mélange

Sélection de variables pour un modèle de mélange fini de régressions,

Bezza Hafidi and Abdallah Mkhadri

Bezza Hafidi and Abdallah Mkhadri On considère dans cette note le problème de sélection de variables pour un modèle de mélange fini de régressions. Récemment, deux approches fondées sur la pénalisation de la log-vraisemblance par deux pénalités convexes de type-Lasso ont été considérées dans la littérature. Mais ces approches présentent certaines limitations méthodologiques ou calculatoires. Nous proposons une nouvelle méthode fondée sur le même principe avec une pénalité composée d'une pénalité L_1-Fusion et d'une autre de type Concave Minimaxe (MC+). Un algorithme type EM est proposé dont les solutions sont obtenues de manière explicite. De plus, les coefficients sont calculés de manière séquentielle coefficient par coefficient. Une étude empirique est réalisée sur des données simulées pour comparer la performance des trois méthodes.

Clustering et visualisation dans le sous-espace discriminant de Fisher : quelques avancées récentes,

Camille Brunet and Charles Bouveyron

Le clustering et la visualisation de données de grandes dimensions sont deux enjeux statistiques récurrents et actuels qui sont confrontés à des problèmes calculatoires ainsi qu'à des difficultés d'interprétation. Récemment, une nouvelle méthode de clustering pour les données de grandes dimensions, appelée Fisher-EM a été proposée par Bouveyron et Brunet (2010) et permet simultanément de classer et de visualiser des données. Cette approche se base sur une modélisation des clusters par modèle de mélanges dans un sous-espace latent discriminant de petite dimension. Bien qu'une telle approche présente de très bonnes performances de clustering aux vues des méthodes existantes et fournit une visualisation informative de l'agencement des données, l'interprétation des clusters obtenus reste limitée puisque les axes discriminants résultent d'une combinaison linéaire des variables d'origine. Aussi, dans le but de faciliter l'interprétation des résultats du clustering, nous proposons une méthode de sélection de variables discriminantes qui, au moyen d'une pénalisation type l1, va introduire de la parcimonie dans l'estimation des axes discriminants de l'algorithme Fisher-EM.

Inférence dans le Stochastic Block Model pour de grands graphes,

Antoine Channarond, Jean-Jacques Daudin and Stéphane Robin

Le modèle SBM est un modèle de mélange pour les graphes aléatoires. L'inférence dans ce modèle est un défi dans les grands graphes de l'ordre de quelques millions de noeuds. Cet article propose une méthode consistante et efficace pour cette taille de graphe, basée sur une classification préalable. La concentration rapide des degrés dans ce modèle permet en effet de classer sans erreur les noeuds à condition que le graphe soit assez grand.

Label switching dans les mélanges, *Christophe Biernacki and Vincent Vandewalle*

L'ensemble logiciel MIXMOD (MIXture MODelling) permet de traiter des problématiques de classification supervisée et non supervisée de données quantitatives ou qualitatives dans un contexte de modèle de mélange. Différents algorithmes d'estimation des paramètres du mélange sont proposés (EM, CEM, SEM) et il est possible de les combiner pour obtenir des stratégies susceptibles de fournir un optimum pertinent de la vraisemblance observée ou complétée. Plusieurs critères d'information pour choisir un modèle parcimonieux (le nombre de composants du mélange notamment) sont disponibles. En plus des mélanges gaussiens multivariés pour traiter les données quantitatives et des mélanges multinomiaux multivariés pour les données catégorielles, MIXMOD propose depuis peu des modèles spécifiques pour traiter les données de grande dimension. Disponibles dans le cadre supervisé depuis 2 ans, ils le seront également dans le cadre non supervisé au cours de l'année 2011. MIXMOD se compose d'une bibliothèque de calcul robuste et performante et d'outils complémentaires : des fonctions pour Matlab et une interface graphique (mixmodGUI).

Inférence bayésienne sur un modèle de mélange à interaction spatiale, *Lionel Cucala and Jean-Michel Marin*

Nous présentons un algorithme MCMC permettant d'estimer les paramètres d'un modèle de mélange avec interaction spatiale. La difficulté provient de la présence d'une constante de normalisation non calculable : nous nous en affranchissons en utilisant la méthode de Murray *et al.*(2006) consistant à générer des variables aléatoires auxiliaires. Ensuite, nous proposons une technique de sélection du nombre de composantes du mélange basée sur l'approximation de Chib (1995). Nous illustrons ces techniques par l'analyse d'images satellites.

Sélection de variables pour un modèle fini de mélange de régressions

Bezza Hafidi [a] & Abdallah Mkhadri [b]

[a] *Faculty of science, University Ibn Zohr, Agadir, Morocco- hbezza@yahoo.fr*
[b] *Faculty of science semlalia, University Cadi Ayyad, Marrakech, Morocco- mkhadri@uacm.ac.ma*

Résumé

On considère dans cette note le problème de sélection de variables pour un modèle de mélange fini de régressions. Récemment, deux approches fondées sur la pénalisation de la log-vraisemblance par deux pénalités convexes de type-Lasso ont été considérées dans la littérature. Mais ces approches présentent certaines limitations méthodologiques ou calculatoires. Nous proposons une nouvelle méthode fondée sur le même principe avec une pénalité composée d'une pénalité L_1-Fusion et d'une autre de type Concave Minimaxe (MC+). Un algorithme type EM est proposé dont les solutions sont obtenues de manière explicite. De plus, les coefficients sont calculés de manière séquentielle coefficient par coefficient. Une étude empirique est réalisée sur des données simulées pour comparer la performance des trois méthodes.

keywords Sélection de variables, Lasso, mélange de régressions, algorithme EM.

Abstract

In this note, we consider the problem of variable selection in finite mixture regression. Recently, in the literature, two approaches based on the penalized likelihood with convex functions type Lasso, have been proposed. However, they had some limitations and they are computationally expensive. We proposed a new approach using a combination of the L_1-Fusion and MC+ penalties. An EM-algorithm for efficient numerical computations is proposed. The performance of our approach is studied in some examples of simulations.

keywords Variable selection , Lasso, mixture regression, EM algorithme.

1 Introduction

Les modèles de mélange fini de régression forment une famille de modèles intéressante et flexible pour modéliser de grandes populations hétérogènes. Ils ont été utilisés avec succès dans de nombreux domaines : astronomie, biologie, les sciences sociales, classification,...,etc. Ils permettent de classer les données en groupe et d'estimer les paramètres du modèle étudié (voir McLachlan & Peel 2000, Tittrington et al 1985).

La sélection de variables est un problème fondamental dans la modélisation statistique qui avait attiré beaucoup d'attention dans les dernières années. Bien que beaucoup de méthodes classiques de sélection de modèle peuvent être utilisées, comme le critère d'information de Akaike (Akaike, 1973) et le critère d'information de Bayes (BIC Schwarz, 1978), néanmoins elles sont peu choisies en pratique à cause de leur complexité algorithmique et leur variabilité, surtout lorsque le nombre de variables observées est assez grand (Brieman, 1996).

Dans le cadre de régression linéaire, Tibshirani(1996) a proposé un critère, appelé LASSO, fondé sur la minimisation des moindres carrés pénalisés pour la sélection de variables. Le LASSO (Least absolute shrinkage and selection operator) consiste à rétrécir et sélectionner les variables les plus pertinentes. Fan & Li (2001,2002) ont proposé aussi le SCAD (Smoothly clipped absolute deviation). Récemment Kalili & Chen (2007) ont développé une méthode fondée sur la maximisation de la vraisemblance pénalisée pour la sélection de variables d'un mélange fini de régression. Ils ont considéré trois type de pénalité ; L_1 (LASSO), HARD et SCAD. Leur méthode, appelé MR-ALasso, est consistante et elle estime et sélectionne les variables simultanément via un algorithme type EM relativement compliqué. La fonction à maximiser dans les méthodes citées sont convexes et ceci en utilisant des méthodes de l'approximation quadratique (cf. Fan and Li 2001). De plus, la pénalité utilisée rétrécie les coefficients de régression de manière individuelle pour chaque classe, et elle n'utilise pas l'information que β_{kj} et $\beta_{k'j}$ sont associés à la même variable j et doivent être traités différemment de $\beta_{kj'}$ qui associée à une différente variable j'. Par ailleurs et dans le même cadre, Luo et al. (2008) ont proposé une combinaison de deux pénalités de type Lasso appelé MR-Lasso dont l'une tienne compte de cette dernière limitation de MR-ALasso. Mais l'algorithme proposé n'est pas nécessairement de type EM. Dans ce travail, nous allons présenter une nouvelle méthode, appelée MRCM-Lasso, pour la sélection de variables d'un mélange de régression en pénalisant la vraisemblance par deux types de pénalités. La première est la pénalité fusion, similaire à celle utilisée par MR-Lasso, qui nous permet de déterminer l'importance de chaque variable au sein de chaque classe et qui a été utilisé par Guo et al (2009) en classification basée sur le mélange de modèles Gaussien. La deuxième pénalité est concave et appelée MC+ par Zhang (2010). L'intérêt de cette dernière, réside dans le fait qu'on n'a pas besoin d'un estimateur initial des coefficients de régression comme dans le cas de la pénalité L_1 adaptative.

Nous nous concentrons dans cette note sur les aspects d'estimation et sélection de variables via des algorithmes type EM. Nous redéfinissons correctement l'algorithme EM pour MR-Lasso. Nous proposons en suite un algorithme type EM pour MRCM-Lasso dont les solutions sont obtenues de manière explicite. De plus, les coefficients sont calculés de manière séquentielle coefficient par coefficient. Une étude empirique est réalisée sur des données simulées pour comparer la performance des trois méthodes.

Bibliographie

[1] Akaike.H (1973), Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csaki, Eds.. *Second International Symposiumon Information Theory*, pp. 267-281.
[2] Breiman, L. (1996), Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**,2350-2383.
[3] Fan, J., and Li, R. (2001), Variable Selection via Non-Concave Penalized Likelihood and Its Oracle Properties, *Journal of the American Statistical Association*, 96, 1348-1360.
[4] Fan, J., and Li, R. (2002), Variable Selection for Cox's Proportional Hazards Model and Frailty Model, *The Annals of Statistics*, 30, 74-99. McLachlan, G. J., and Peel, D. (2000), Finite Mixture Models, New York: Wiley.
[5]Guo F.J, Levina E., Michailidis G., Zhu J. (2009) Pairwise Variable Selection for High-dimensional Model-based Clustering. To appear in Biometrics.
[6] Luo, R., Wang, H. and Tsai, C.-L. (2008). On mixture regression shrinkage and selection via the MR-Lasso. *International Journal of Pure and Applied Mathematics*, 403-414.
[7] Khalili, A., and Chen, J. (2007), Variable Selection in Finite Mixture of Regression Models, *Journal of the American Statistical Association*,
[8] Schwarz. G (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.
[9] Tibshirani, R. (1996), Regression Shrinkage and Selection via the LASSO, *Journal of the Royal Statistical Society*, Ser. B, 58, 267-288.
[10] Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985), Statistical Analysis of Finite Mixture Distributions, New York: Wiley.
[11] Zhang, C. (2010), Nearly Unbiased Variable Selection under Minimax Concave Penalty, *The Annals of Statistics*, Vol. 38, No. 2, 894-942.

Clustering et visualisation dans le sous-espace discriminant de Fisher: quelques avancées récentes.

Camille Brunet[1] & Charles Bouveyron[2]

1 *Université d'Evry - Laboratoire IBISC, 40 rue de Pelvoux CE1455, 91020 Evry Courcouronnes.*

2 *Université Paris 1-Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris.*

Résumé: Le clustering et la visualisation de données de grandes dimensions sont deux enjeux statistiques récurrents et actuels qui sont confrontés à des problèmes calculatoires ainsi qu'à des difficultés d'interprétation. Récemment, une nouvelle méthode de clustering pour les données de grandes dimensions, appelée Fisher-EM a été proposée par [1] et permet simultanément de classer et de visualiser des données. Cette approche se base sur une modélisation des clusters par modèle de mélange gaussien dans un sous-espace latent discriminant de petite dimension. Bien qu'une telle approche présente de très bonnes performances de clustering aux vues des méthodes existantes et fournit une visualisation informative de l'agencement des données, l'interprétation des clusters obtenus reste limitée puisque les axes discriminants résultent d'une combinaison linéaire des variables d'origine. Aussi, dans le but de faciliter l'interprétation des résultats du clustering, nous proposons une méthode de sélection de variables discriminantes qui, au moyen d'une pénalisation type L1, va introduire de la parcimonie dans l'estimation des axes discriminants de l'algorithme Fisher-EM.

Mots clefs: classification non-supervisée en grande dimension, modèle de mélanges, sous-espace discriminant, critère de Fisher, visualisation, modèle parcimonieux, pénalisation ℓ_1, sélection de variables.

Abstract: The clustering and the visualization of high-dimensional data is a recurrent and challenging task which poses computational problems and difficulties for result interpretation. Recently, a new clustering method for high-dimensional data, called Fisher-EM, has been proposed by [1] which simultaneously clusters and visualizes the data. This approach models the clusters by mixture of Gaussians in a latent discriminative subspace of low dimension. Admittedly, the clustering accuracy of the Fisher-EM algorithm outperforms most of existing methods, nevertheless the provided clustering results remain difficult to understand since the discriminative axes are a linear combination of original variables. In order to ease the interpretation of the clustering, this work proposes a method for variable selection by introducing sparsity through a ℓ_1 penalization into the discriminative axes of the Fisher-EM algorithm.

Key words: high-dimensional clustering, model-based clustering, discriminative subspace, Fisher criterion, visualization, parsimonious models, ℓ_1 penalization, variable selection.

Résumé long

Le clustering et la visualisation de données de grandes dimensions sont deux enjeux statistiques récurrents et actuels qui sont confrontés à des problèmes calculatoires ainsi qu'à des difficultés d'interprétation. Certes, il existe en clustering des approches combinant la réduction de dimension et la classification non supervisée qui sont basées sur la recherche de sous-espaces propres aux classes, en particulier les travaux de [4, 2, 5]. Cependant, malgré les très bonnes performances de ces nouvelles approches, une visualisation informative des clusters résultants de celles-ci n'est pas possible. Pour faire face à de tels problèmes, nous avons proposé une méthode probabiliste et un algorithme appelé Fisher-EM, qui permet simultanément de classer et de visualiser des données dans un contexte de classification non supervisée. Cette approche se base sur une modélisation des clusters par modèles de mélange, dans un sous-espace latent discriminant de petite dimension estimé par l'intermédiaire d'un critère basé sur la théorie de Fisher [3].

Considérons $\{y_i\}_{i=1}^n \in \mathbb{R}^p$ les données à classer en K groupes homogènes *i.e.* adjoindre une valeur $z_i = \{1, \ldots, K\}$ à l'observation y_i où $z_i = k$ indique l'appatenance de l'observation y_i au groupe k. De plus, on fait l'hypothèse que $\{y_i\}_{i=1}^n$ et $\{z_i\}_{i=1}^n$ sont des réalisations indépendantes de vecteurs aléatoires $Y \in \mathbb{R}^p$ et $Z \in \{1, \ldots, K\}$ respectivement. Par ailleurs, on définit $\mathbb{E} \subset \mathbb{R}^p$ un espace latent de dimension $d \leq K-1$, représentant l'espace le plus discriminant tel que $0 \in \mathbb{E}$ et de dimension d strictement plus petite que la dimension p de l'espace des observations. Dans cet espace latent, $\{x_i\}_{i=1}^n \in \mathbb{E}$ représentent les données réelles qui sont des réalisations indépendantes d'un vecteur aléatoire $X \in \mathbb{E}$. Le lien existant entre la variable observée $Y \in \mathbb{R}^p$ et celle de l'espace latent $X \in \mathbb{E}$ est supposé être une transformation linéaire,

$$Y = UX + \varepsilon, \tag{1}$$

où U est une matrice de taille $p \times d$ vérifiant $U^t U = \mathbf{I}_d$ et ε est un terme de bruit. On suppose de plus que,

$$X|_{Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k) \text{ et } \varepsilon|_{Z=k} \sim \mathcal{N}(0, \Psi_k) \tag{2}$$

où $\mu_k \in \mathbb{E}$ et $\Sigma_k \in \mathbb{R}^{d \times d}$ représentent respectivement le vecteur moyennes et la matrice de covariances de la classe k dans l'espace latent $\mathbb{E}$ et $\Psi_k \in \mathbb{R}^{p \times p}$ la matrice de covariances associée au bruit. D'après l'équation 1, ces hypothèses impliquent que la distribution de

probabilié conditionnelle de Y est $Y|_{X,Z=k} \sim \mathcal{N}(UX, \Psi_k)$ et il en résulte que sa distribution marginale est un mélange de gaussiennes:

$$f(y) = \sum_{k=1}^{K} \pi_k \phi(y; m_k, S_k)$$

où π_k, $m_k = U\mu_k$ et $S_k = U^t \Sigma_k U + \Psi_k$ représentent respectivement la proportion, le vecteur moyenne et la matrice de covariances de la classe k de l'espace des observations et $\phi(.)$ une densité gaussienne. On introduit $W = [U, V]$ une matrice $p \times p$ qui vérifie $W^t W = WW^t = \mathbf{I_p}$ où $V \in \mathbb{R}^{p \times (p-d)}$ est le complément orthogonal de U définie précédemment. Enfin, on suppose que la variance du bruit Ψ_k vérifie $V^t \Psi_k V = \beta_k \mathbf{I_p}$ et $U^t \Psi_k U = \mathbf{O_d}$ telle que $\Delta_k = W^t S_k W = \text{diag}(\Sigma_k, \beta_k \mathbf{I_{p-d}})$.
On fera référence à ce modèle comme étant le modèle $\text{DLM}_{[\Sigma_k \beta_k]}$. Notons qu'en imposant des contraintes sur les matrices de covariances Σ_k et $\beta_k \mathbf{I_{p-d}}$, il est possible de décliner 11 autres modèles (cf. [1]).

En considèrant l'ensemble de ces hypothèses, l'espérance de la log-vraisemblance compléte $Q(y_1, \ldots, y_n; \theta)$ du modèle $\text{DLM}_{[\Sigma_k \beta_k]}$ s'écrit:

$$\begin{aligned} Q(\theta) &= -\frac{1}{2} \sum_{k=1}^{K} n_k [-2 \log(\pi_k) + \text{trace}\left(\Sigma_k^{-1} U^t C_k U\right) + \log(|\Sigma_k|) \\ &\quad + (p-d) \log(\beta_k) + \frac{1}{\beta_k} \left(\text{trace}(C_k) - \sum_{j=1}^{d} u_j^t C_k u_j \right) + \gamma]. \end{aligned} \tag{3}$$

où C_k est la matrice de covariances empirique de la classe k, u_j est le $j$ème vecteur colonne de la matrice U, $n_k = \sum_{i=1}^{n} t_{ik}$ où t_{ik} est la probabilité a posteriori qu'une observation y_i appartienne à la classe k et $\gamma = p \log(2\pi)$ est une constante.

Dans un contexte non supervisé la maximisation directe de $Q(\theta)$ n'étant pas faisable, nous avons donc utilisé une procédure itérative en trois étapes: une étape E qui calcule à chaque itération (q), l'espérance de la log vraisemblance compléte conditionnellement à la valeur courante du paramètre $\theta^{(q-1)}$. En pratique, cela se résume au calcul de la probabilité a posteriori $t_{ik}^{(q)} = E[z_{ik}|y_i, \theta^{(q-1)}]$ qu'une observation y_i appartienne au $k$ème cluster; une étape F qui détermine la transformation linéaire $U \in \mathbb{R}^{p \times d}$, sous contrainte d'orthogonalité de ses colonnes, relative à la base du sous-espace latent de dimension $d = K - 1$ dans lequel les K groupes sont le mieux séparés. Pour cela, nous avons adapté le problème de maximisation du traditionnel critère de Fisher $J(U) = tr((U^t S U)^{-1} U^t S_B U)$, habituellement utilisé dans un contexte supervisé, à un contexte non supervisé sous la contrainte d'orthogonalité et conditionnellement à la partition floue courante des données. Enfin une étape M estime les paramètres du $\text{DLM}_{[\alpha_{k_j} \beta_k]}$ modèle en maximisant l'espérance conditionnelle de la log-vraisemblance compléte $Q(y_1, \ldots, y_n, \theta)$ définie par l'expression 3.

Bien qu'une telle approche présente de très bonnes performances de clustering aux vues des méthodes existantes et fournit une visualisation informative de l'agencement des données, il n'en reste pas moins que l'interprétation des clusters obtenus reste limitée puisque les axes discriminants résultent d'une combinaison linéaire des variables d'origine. Aussi, dans le but de faciliter l'interprétation et la visualisation des résultats du clustering, nous vous proposerons une méthode de sélection de variables discriminantes qui, au moyen d'une pénalisation type $\ell 1$, va introduire de la parcimonie dans l'estimation des axes discriminants de l'algorithme Fisher-EM.

References

[1] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Preprint*, HAL 00492406:1–44, 2010.

[2] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Communication in Statistics: Theory and Methods*, 52(1):502–519, 2007.

[3] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[4] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, (41):379, 2003.

[5] P. McNicholas and B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

Inférence dans le Stochastic Block Model pour de grands graphes

Antoine Channarond & Jean-Jacques Daudin & Stéphane Robin

UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard 75005 Paris

Résumé

Le modèle SBM est un modèle de mélange pour les graphes aléatoires. L'inférence dans ce modèle est un défi dans les grands graphes de l'ordre de quelques millions de noeuds. Cet article propose une méthode consistante et efficace pour cette taille de graphe, basée sur une classification préalable. La concentration rapide des degrés dans ce modèle permet en effet de classer sans erreur les noeuds à condition que le graphe soit assez grand.

Abstract

The inference in the SBM model is a real challenge in large graphs, of millions of nodes. This article gives a new method that is consistent and efficient in such graphs, based on a preliminary classification. Thanks to the fast concentration of the degrees in this model, it is possible to classify without any mistake, as soon as the graph is large enough.

1 Stochastic Block Model

1.1 Présentation du modèle

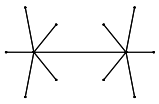
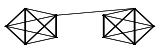
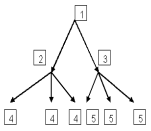
Le modèle classique d'Erdős-Rényi n'est pas satisfaisant pour modéliser la plupart des graphes issus de données réelles, comme les réseaux d'interaction de protéines ou les réseaux sociaux, à cause de l'uniformité du comportement de connexion parmi les sommets.

L'idée du Stochastic Block Model (SBM) est de rendre compte de la structure (sociale par exemple) dans la population des sommets en attribuant une classe d'appartenance à chaque sommet, puis en tirant les arêtes conditionnellement aux classes ; cf Holland, Laskey et Leinhardt (1983). Ainsi SBM est un modèle relativement souple permettant de modéliser des graphes aux topologies très typées comme les communautés, les étoiles ou les hiérarchies par exemple, voir le tableau.

Le modèle se définit formellement de la façon suivante. Soit $n \geq 1$. On note l'ensemble des sommets $V_n = \{1, \ldots, n\}$, et on désigne un graphe à n sommets par sa matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$. Soit $Q \geq 1$ le nombre de classes, qui est supposé connu.

– Soit $(Z_i)_{1 \leq i \leq n}$ une suite de variables indépendantes et identiquement distribuées de loi multinomiale sur $\{1, \ldots, Q\}$ de paramètre $\alpha = (\alpha_1, \ldots, \alpha_Q)$.

TABLE 1 – Exemples de modèles SBM.

Description	Graphe	Q	π
Étoiles		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Communautés		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$
Hiérarchie		5	$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

– Soit (X_{ij}) une matrice de variables indépendantes conditionnellement à Z. Pour tout $i, j \in V_n$, et $q, r \in \{1, \ldots, Q\}$, conditionnellement à $Z_i = q, Z_j = r$, X_{ij} suit une loi de Bernoulli de paramètre π_{qr}.

Du point de vue de la modélisation, seules les arêtes sont observées, tandis que la structure de la population reste inconnue. Les variables (Z_i) seront donc dites cachées, et ne pourront pas être utilisées pour l'inférence, qui devient dès lors un défi dans les graphes de grande taille.

1.2 Enjeux de l'inférence du modèle

La complexité du graphe de dépendance des classes conditionnellement aux arêtes est telle que l'utilisation directe du maximum de vraisemblance ou d'un algorithme EM simple est impossible. Nowicki et Snijders (2001) mettent en oeuvre une méthode MCMC, et Daudin, Picard et Robin (2008) une méthode d'inférence approchée, dite variationnelle [1], l'une ne permettant de traiter des graphes que de quelques centaines de noeuds, l'autre quelques milliers en temps raisonnable. La méthode présentée dans la suite traite des graphes de plusieurs millions de noeuds.

Les méthodes évoquées traitent conjointement la classification des sommets et l'estimation des paramètres. Bien qu'apparaissant comme un sous-produit de l'inférence dans ces méthodes, la classification des sommets en est en fait l'obstacle majeur. Si les classes des sommets étaient dévoilées, l'inférence se résumerait en effet à des estimateurs basiques.

Or justement la structure cachée de ce modèle se dévoile d'autant plus qu'il y a d'individus. Le modèle jouit en effet d'une propriété spécifique aux graphes : quand on y ajoute

1. Voir Wainwright et Jordan (2008) pour les méthodes variationnelles.

un sommet, on ajoute ses informations propres — à qui il se connecte —, mais on ajoute aussi une information à tous les sommets précédents — le fait de s'être connecté ou non à ce nouveau sommet —.

On se propose donc de commencer par établir une classification des sommets, afin de connaître au mieux les classes, puis d'estimer les paramètres grâce à elle. Elle sera basée sur une variable tenant compte de l'information donnée par les nouvelles connexions : les degrés. Il est à noter qu'il existe d'autres méthodes permettant de retrouver les classes du modèle SBM, par exemple le spectral clustering dans Rohe, Chatterjee et Yu (2010).

2 Méthode d'estimation basée sur les degrés

2.1 Concentration des degrés normalisés

On rappelle que le degré D_i d'un sommet $i \in V_n$ est le nombre de ses voisins, soit $D_i = \sum_{j \neq i} X_{ij}$. On appelle degré normalisé du sommet $i \in V_n$: $T_i = \frac{D_i}{n-1}$. La loi du degré d'un sommet conditionnellement à la classe q du sommet est binomiale de paramètres $(n, \overline{\pi}_q)$, où $\overline{\pi}_q = \sum_r \alpha_r \pi_{qr}$, ce qui confère aux degrés normalisés un comportement de concentration rapide autour de leur moyenne, illustré par l'inégalité suivante :

$$\forall q \in \{1, \dots Q\}\, \forall t > 0\; P(|T_i - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2}$$

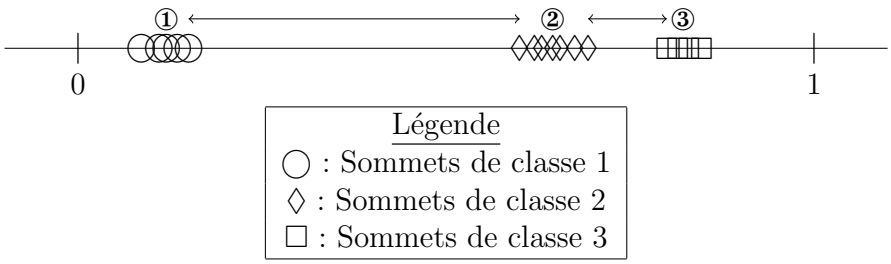
Le phénomène de concentration conduit à la formation d'agrégats de points de même classe, ce qui permet d'identifier facilement les classes de sommets quand n est suffisamment grand.

2.2 Stratégie d'inférence

L'idée est de trier d'abord les sommets selon la proximité de leurs degrés normalisés. L'algorithme proposé consiste à calculer les écarts entre les degrés normalisés consécutifs et à trouver les $Q-1$ plus grands écarts de sorte à former Q groupes de sommets, qui sont normalement les classes cachées. Le schéma ci-après montre la répartition des degrés normalisés pour un modèle avec $Q = 3$ classes. Les deux plus grands écarts sont indiqués par les flèches, et les trois groupes formés par l'algorithme par ①, ② et ③. Ce cas est favorable, puisque les groupes formés correspondent aux vraies classes.

On construit les estimateurs usuels des proportions et des connectivités en substituant la vraie partition des sommets en classes par la partition estimée par la classification. Si on note $(\widehat{\mathcal{C}}_q)_{1 \leq q \leq Q}$ la partition estimée et pour tout $1 \leq q \leq Q$, N_q le cardinal de $\widehat{\mathcal{C}}_q$, ces estimateurs s'écrivent ainsi :

$$\widehat{\alpha}_q = \frac{N_q}{n} \quad \widehat{\pi}_{qr} = \frac{1}{N_q N_r} \sum_{(i,j) \in \widehat{\mathcal{C}}_q \times \widehat{\mathcal{C}}_r} X_{ij}$$



Le paramètre déterminant pour l'efficacité de l'algorithme — et de l'estimation en découlant — est la proximité entre les différentes moyennes des degrés normalisés. Plus elles sont proches, plus n doit être grand pour que les agrégats de points se séparent complètement. On la mesure par le paramètre $\delta = \min_{q \neq r} |\overline{\pi}_q - \overline{\pi}_r|$.

Cette méthode d'estimation est rapide. Au lieu de répéter un grand nombre de fois des étapes de complexité quadratique — mises à jour successives des paramètres puis de la classification —, on effectue seulement deux étapes de complexité en le nombre d'arêtes (moins que quadratique) pour le calcul des degrés, et pour le calcul des estimateurs. A la différence des autres méthodes, la classification et l'estimation sont effectuées séparément.

2.3 Consistance de la méthode

Par l'argument de concentration, on montre que l'algorithme est consistant dès que $\delta > 0$; la probabilité de l'événement "il existe une erreur de classification", noté E, étant majorée ainsi :

Théorème.

$$P(E) \leq 2ne^{-\frac{2}{25}n\delta^2} + Q(1-\alpha_0)^{n+1}$$

où $\alpha_0 = \min_q \alpha_q$. Notons que l'algorithme n'est efficace que si $\delta \gg \sqrt{\frac{\ln n}{n}}$. Dans le cas dégénéré où il existe au moins deux moyennes égales ($\delta = 0$), il est possible d'appliquer la stratégie à une quantité supplémentaire pour séparer des groupes dont les degrés normalisés se concentrent au même endroit. Par suite :

Théorème. $(\widehat{\alpha}, \widehat{\pi})$ *est un estimateur consistant des paramètres* (α, π) *du modèle.*

2.4 Critère de sélection de modèle

Jusqu'à présent, le nombre Q de classes était supposé connu. Afin de rendre cette méthode autosuffisante, on peut aller plus loin dans l'étude du comportement des écarts entre les degrés normalisés pour en tirer un critère de sélection de modèle. Le but en est de détecter le nombre convenable de classes, en détectant le nombre d'agrégats formés par les degrés normalisés, grâce encore une fois à leur comportement de concentration.

Notons $(G_q)_{1\leq q\leq Q-1}$ la suite décroissante des écarts entre degrés normalisés successifs, et $(H_q)_{1\leq q\leq Q-1}$ la suite décroissante des écarts entre moyennes successives des groupes issus de la classification. On note ainsi le critère :

$$f_Q = \sum_{1\leq q\leq Q-1} |G_q - H_q| + \frac{1}{n^{\frac{1-\beta}{2}} G_{Q-1}} \text{ où } 0 < \beta < 1.$$

La première somme est grande quand le nombre de classes est trop petit, car les écarts entre degrés ne correspondent pas aux écarts entre moyennes des groupes qui en englobent en fait plusieurs. En revanche, le second terme est grand lorsque le nombre de classes est trop grand, car certains des plus grands écarts $(G_q)_{1\leq q\leq Q}$ sont situés entre les degrés de sommets de même classe. Quand le nombre de classes est correct, le critère tend vers 0. On démontre suivant ce schéma :

Théorème. $\widehat{Q} = \arg\min_Q f_Q$ *est un estimateur consistant du nombre de classes du modèle.*

2.5 Simulations

En classification, on verra que les simulations révèlent que l'algorithme est efficace comme prévu quand $\delta \gg \sqrt{\frac{\ln n}{n}}$, mais pas avant. Le problème de cet algorithme est qu'il fait confiance de la même façon à tous les degrés normalisés, même isolés et donc sans doute peu représentatifs de la classe. Cela arrive quand la concentration trop faible, en particulier lorsque n est trop petit. Si on veut éliminer les points douteux, on peut utiliser un autre algorithme qui choisit les Q mailles les plus remplies de l'histogramme des degrés normalisés. Cet algorithme est aussi consistant si le pas du maillage est bien réglé. Il est meilleur lorsque n est petit, mais perd l'avantage asymptotiquement.

On verra de plus que l'estimation des paramètres est très bonne dans les conditions prévues, et même assez bonne avant, même si la classification n'est pas encore complètement exempte d'erreurs.

Bibliographie

[1] Daudin, J., F. Picard, et S. Robin (2008). A mixture model for random graphs. Statistics and computing 18(2), 173–183.
[2] Nowicki, K. et T. Snijders (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455), 107–1087.
[3] Wainwright, M. et M. Jordan (2008). Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning 1(1-2), 1–305.
[4] Rohe, K., S. Chatterjee, et B. Yu (2010). Spectral clustering and the high-dimensional Stochastic Block Model. Arxiv preprint arXiv :1007.1684.
[5] P.W. Holland, K.B. Laskey, et S. Leinhardt. Stochastic blockmodels : First steps. Social Networks, 5(2) :109–137, 1983.

Label switching dans les mélanges

Christophe Biernacki[1] & Vincent Vandewalle[2]

[1] *Université Lille 1 & CNRS & INRIA – 59655 Villeneuve d'Ascq Cedex (France)*
[2] *Université Lille 2 & INRIA – 59100 Roubaix (France)*

Résumé :

Nous proposons une loi *a posteriori* dont la partition latente est restreinte à une numérotation particulière conduisant à la plus grande séparation avec ses permutations. Deux mesures de séparation sont proposées, l'une *globale* mais calculable uniquement pour de très petites tailles d'échantillon (divergence de Kullback), l'autre *locale* et donc très facile à calculer (écart des lois en le MAP). Un algorithme de Gibbs permet de simuler facilement suivant cette nouvelle loi. Cette procédure est suffisamment générale pour s'appliquer directement à toutes les distributions et des expériences dans un cas gaussien très simple montrent des résultats particulièrement encourageants.
mots-clés : statistique bayesienne, modèles de mélange, label switching, algorithme de Gibbs

Abstract :

We propose a posterior distribution for which the latent partition is restricted to a special numbering leading to the largest separation with its permutations. Two different measures of separation are proposed, the first one being *global* but untractable even from very small sample sizes (Kullback divergence), the second one being *local* and thus very easy to compute (difference of distributions at the MAP). A Gibbs algorithm allows to sample easily according to this new distribution. This procedure is general enough to apply directly with any distribution and some experiments in a basic Gaussian situation show particular encouraging results.
keywords: Bayesian statistics, mixture models, label switching, Gibbs algorithm

1 Introduction

Au cours des quinze dernières années il y a eu un très grand intérêt autour de l'utilisation des méthodes bayésiennes dans les mélanges de distributions. Une des raisons de ce succès est l'émergence des méthodes MCMC. Cependant, un des principaux problèmes de ces méthodes est la non-identifiabilité des composants causée par des *a priori* symétriques, ce qui conduit à l'inutilité des sorties de l'algorithme de Gibbs pour l'inférence ; ce problème est connu sous le nom de label switching. Jasra *et al.* (2005) ont dressé un état de l'art complet des solutions apportées à ce problème et nous en rappelons ici les quatre principales.

Une première solution consiste à imposer des contraintes d'identifiabilité artificielles sur les paramètres (Diebolt et Robert, 1994), comme des contraintes d'ordre ($\theta_1 < \theta_2$ par

exemple) pouvant être interprétées comme une modification de la loi *a priori*. En pratique la rupture de la symétrie au niveau de la loi *a priori* ne suffit pas à régler le problème (Celeux et al., 2000 ; Jasra *et al.*, 2005).

Une seconde solution consiste à utiliser des algorithmes de réétiquetage des paramètres générés (Stephens, 1997 ; Celeux, 1998). Il s'agit pour un paramètre généré par l'échantillonneur de Gibbs de trouver sa permutation qui minimise une fonction de perte. La mise en œuvre de cet algorithme conduit à un algorithme de type k-means sur l'espace des paramètres. Cette approche souffre des mêmes défauts que ce dernier quand les composants sont peu séparés (Celeux, 1997), ce qui conduit à une sous-estimation de la variabilité de la loi *a posteriori*.

Une troisième solution consiste à utiliser des fonctions de perte invariantes à la permutation des paramètres (Celeux et *al.*, 2000). L'utilisation de cette stratégie nécessite le choix d'une fonction de perte adaptée au problème inférentiel, ainsi que l'optimisation de cette dernière.

Enfin, puisque les méthodes précédentes ne tiennent pas compte de l'incertitude d'attribution de la permutation aux paramètres, une approche probabiliste a été proposée (Jasra et al., 2005 ; Sperrin et al., 2010). Il s'agit de construire un modèle sur la distribution *a posteriori* non switchée. En pratique, une séquence de paramètres pour laquelle le switch n'est pas encore intervenu est utilisée pour construire un modèle sur la distribution *a posteriori* non switchée. Cette quantité est ensuite utilisée pour attribuer une probabilité à chaque permutation du paramètre obtenu par l'échantillonneur de Gibbs. Cette probabilité peut alors être utilisée pour le calcul de quantités d'intérêt telles que la moyenne *a posteriori*.

La méthode probabiliste est une avancée intéressante pour le problème du label switching, cependant elle nécessite d'une part un réglage visuel pour déterminer si la distribution n'a pas switchée lors de la construction du modèle, et d'autre part elle nécessite de faire des hypothèses assez fortes sur la distribution *a posteriori* non switchée. Nous proposons maintenant une loi *a posteriori* dont la partition latente est restreinte à une numérotation particulière conduisant à la plus grande séparation avec ses permutations, ainsi qu'un algorithme de Gibbs permettant de simuler facilement suivant cette nouvelle loi. Les deux principaux avantages de la méthode proposée, outre sa réelle simplicité de mise en œuvre, sont alors la prise en compte de l'incertitude d'attribution de la permutation aux paramètres et d'autre part la non nécessité d'imposer un modèle rigide sur la distribution *a posteriori* non switchée.

2 Loi *a posteriori* restreinte par la partition

2.1 Rappel du problème de *label switching*

Nous nous plaçons dans le cadre générique d'un mélange de g lois vérifiant les conditions classiques de régularité

$$p(\cdot|\theta) = \sum_{k=1}^{g} \alpha_k p(\cdot|\beta_k)$$

où les α_k correspondent aux proportions du mélange ($\alpha_k > 0$ et $\sum_k \alpha_k = 1$) et les β_k sont les paramètres des différentes lois du mélange. On regroupe aussi l'ensemble des paramètres $\theta_k = (\alpha_k, \beta_k)$ en un paramètre global $\theta = (\theta_1, \ldots, \theta_g) \in \Theta$.

Partant d'un n échantillon i.i.d. $x = (x_1, \ldots, x_n)$ issu de $p(\cdot|\theta)$ et d'une loi *a priori* $p(\theta)$, toute inférence bayésienne s'appuie sur la loi *a posteriori*

$$p(\theta|x) \propto p(x|\theta)p(\theta).$$

On remarque alors que $p(\theta|x)$ est invariante à une renumérotation des composantes du mélange dès que $p(x|\theta)$ et $p(\theta)$ le sont aussi. En d'autres termes, notant $\mathcal{P}_g$ l'ensemble des permutations de $\{1, \ldots, g\}$ et $\sigma(\theta) = (\theta_{\sigma(1)}, \ldots, \theta_{\sigma(g)})$ le paramètre θ permuté en indices avec $\sigma \in \mathcal{P}_g$, on a $p(\theta|x) = p(\sigma(\theta)|x)$ pour tout $\sigma \in \mathcal{P}_g$. Cette symétrie exacte de la loi *a posteriori*, appelé aussi problème du *label switching*, rend alors vide de sens le calcul direct de nombreux estimateurs ponctuels habituels comme la moyenne *a posteriori*.

2.2 Définition d'une nouvelle loi *a posteriori*

Afin de pallier ce problème du *label switching*, nous proposons de conditionner la loi *a posteriori* à une numérotation particulière non pas sur le paramètre θ comme cela est fait classiquement mais plutôt sur la partition latente. Pour rappel, le modèle de mélange peut être vu sous la forme générative suivante : chaque x_i provient de la loi $p(\cdot|\beta_{z_i})$ où $z_i \in \{1, \ldots, g\}$ provient lui-même d'une loi multinomiale de paramètre $(\alpha_1, \ldots, \alpha_g)$. On note alors $z = (z_1, \ldots, z_n) \in \mathcal{Z}$ la partition latente qui a servi à générer x.

Notons $\tilde{\mathcal{Z}} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_{g!}\}$ une partition de l'ensemble des partitions $\mathcal{Z}$ qui fixe une numérotation particulière pour chaque partition z de l'ensemble des données x, c'est-à-dire

$$\forall h, h' \in \{1, \ldots, g!\}, \exists! \sigma \in \mathcal{P}_g \text{ tq } z \in \mathcal{Z}_h \Leftrightarrow \sigma(z) \in \mathcal{Z}_{h'}$$

avec $\sigma(z) = (\sigma(z_1), \ldots, \sigma(z_n))$ indiquant que z est permuté en indices pour $\sigma \in \mathcal{P}_g$. Dans ce cadre, la loi *a posteriori* habituelle se décompose comme un mélange de $g!$ lois *a posteriori* conditionnées par n'importe qu'elle numérotation particulière $\tilde{\mathcal{Z}}$ des partitions

$$p(\theta|x) = \sum_{h=1}^{g!} p(\theta|x, \mathcal{Z}_h) p(\mathcal{Z}_h|x) = \frac{1}{g!} \sum_{h=1}^{g!} p(\theta|x, \mathcal{Z}_h).$$

Remarquons que, contrairement à $p(\theta|x)$, l'ensemble des lois $p(\theta|x, \mathcal{Z}_h)$ ne sont plus strictement invariantes à la numérotation de z mais l'importance de cette asymétrie dépend bien entendu du choix de $\tilde{\mathcal{Z}}$. Afin de s'éloigner le plus possible de la symétrie, l'idée clé de notre proposition est alors de choisir un découpage $\tilde{\mathcal{Z}}$ qui sépare au mieux les lois $p(\theta|x, \mathcal{Z}_h)$ de ce mélange puis de retenir comme nouvelle définition de loi *a posteriori* n'importe laquelle de ces $g!$ lois $p(\theta|x, \mathcal{Z}_h)$, par exemple $p(\theta|x, \mathcal{Z}_1)$, le choix d'un h particulier étant bien entendu arbitraire et sans conséquence.

Un premier choix naturel de $\tilde{\mathcal{Z}}$, noté $\tilde{\mathcal{Z}}^{KL}$, est celui qui conduit au plus grand écart de Kullback-Leibler entre les composantes de mélange sur $\mathcal{Z}_h$, ce qui s'écrit

$$\tilde{\mathcal{Z}}^{KL} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \int_{\Theta} p(\theta|x, \mathcal{Z}_1) \ln \left[\frac{p(\theta|x, \mathcal{Z}_1)}{p(\theta|x, \mathcal{Z}_h)} \right] d\theta.$$

Ce critère est inaccessible dès de très petites tailles d'échantillon à cause de la combinatoire sur le nombre de partitions, c'est pourquoi nous proposons aussi un critère plus simple, optimisant l'écart entre les lois en un θ particulier au lieu de l'espace Θ dans son entier. Nous retenons pour cela l'estimateur du maximum *a posteriori* θ^{MAP}, ce qui donne une nouvelle numérotation optimale notée $\tilde{\mathcal{Z}}^{MAP}$ définie par

$$\tilde{\mathcal{Z}}^{MAP} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \frac{p(\theta^{MAP}|x, \mathcal{Z}_1)}{p(\theta^{MAP}|x, \mathcal{Z}_h)}.$$

En pratique, $\tilde{\mathcal{Z}}^{MAP}$ est immédiate à calculer pour toute taille d'échantillon puisque cela revient à prendre la numérotation la plus probable individu par individu et calculée en θ^{MAP} :

$$\mathcal{Z}_1^{MAP} = \left\{ z \in \mathcal{Z} / Id = \arg\max_{\sigma \in \mathcal{P}_g} p(\sigma(z)|x, \theta^{MAP}) \right\},$$

où Id correspond à la permutation identité. On peut interpréter θ^{MAP} comme un paramètre de référence pour la numérotation de la partition latente.

2.3 Simulation suivant un algorithme de Gibbs

L'algorithme de Gibbs classique est légèrement modifié pour tenir compte du conditionnement sur $\mathcal{Z}_1$ mais reste très simple. On tire z suivant la loi habituelle $p(z|x, \theta)$, on permute ensuite z de telle sorte que $\sigma(z) \in {\mathcal{Z}_1}^{KL}$ ou $\sigma(z) \in {\mathcal{Z}_1}^{MAP}$ (suivant le critère retenu), enfin on génère θ suivant $p(\theta|x, \sigma(z))$.

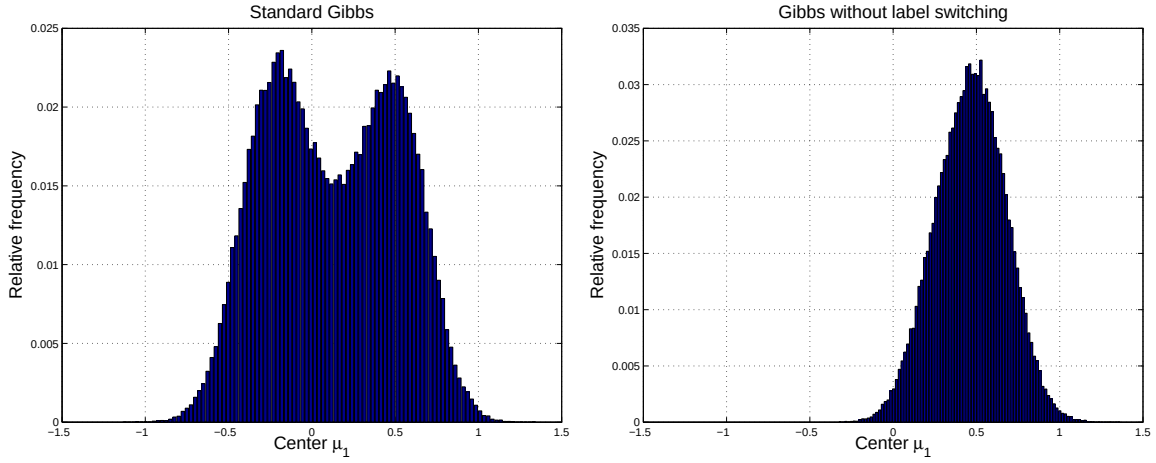
3 Expériences dans le cas gaussien

Afin d'illustrer numériquement l'intérêt de notre proposition, nous prenons un mélange très simple de deux composantes gaussiennes ($g = 2$) univariées $p(\cdot|\beta_k) = \mathcal{N}(\beta_k, 1)$ avec

proportions $\alpha_1 = \alpha_2 = 0.5$. Les proportions et les variances sont donc connues et fixées, seules les moyennes $\theta_1 = \beta_1$ et $\theta_2 = \beta_2$ sont inconnues. On prend par ailleurs la loi *a priori* sur $\theta_k \sim \mathcal{N}(0,1)$ avec $\theta_1 \perp \theta_2$. On obtient donc au final les lois *a posteriori* $\theta_k|z,x \sim \mathcal{N}(n_k\bar{x}_k/(n_k+1), 1/(n_k+1))$ et $z_i|\theta,x \sim \mathcal{M}_2(1, t_{i1}(\theta), t_{i2}(\theta))$ en utilisant les notations classiques $n_k = \sum_{i=1}^n \mathbb{I}_{z_i=k}$, $\bar{x}_k = \sum_{i=1}^n \mathbb{I}_{z_i=k} x_i/n_k$ et $t_{ik}(\theta) = p(z_i = k|x,\theta)$. Dans la suite, on prend $\theta_1 = 0$ et $\theta_2 = 0.25$.

3.1 Visualisation sur un exemple

On génère un échantillon x de taille $n = 100$ et la séquence de chauffe est de 100 itérations pour 100 000 itérations en tout. Les figures 1 (a) et (b) donnent respectivement les lois *a posteriori* $p(\theta|x)$ et $p(\theta|x, \mathcal{Z}_1^{MAP})$. La nouvelle procédure semble bien ne retenir qu'un unique mode pertinent à première vue malgré la forte imbrication des modes de la loi *a posteriori* habituelle.



(a) (b)

Figure 1: (a) : loi *a posteriori* habituelle $p(\theta|x)$, (b) nouvelle loi *a posteriori* proposée $p(\theta|x, \mathcal{Z}_1^{MAP})$.

3.2 Qualité de la moyenne *a posteriori*

On génère cette fois 100 échantillons x de taille $n \in \{3, 10, 100\}$, la séquence de chauffe est maintenue à 100 itérations et on effectue 2000 itérations de Gibbs. Le tableau 1 donne la moyenne de l'erreur quadratique de la moyenne *a posteriori* pour la stratégie habituelle basée sur un Gibbs renuméroté par les k-means et pour les deux nouvelles stratégies proposées $p(\theta|x, \mathcal{Z}_1^{KL})$ et $p(\theta|x, \mathcal{Z}_1^{MAP})$. La loi $p(\theta|x, \mathcal{Z}_1^{KL})$ est uniquement disponible pour $n = 3$ pour des raisons combinatoires. On remarque que les nouvelles procédures

améliorent uniformément et assez significativement (surtout pour n petit) la qualité de l'estimateur en comparaison d'un Gibbs/k-means standard.

Stratégie	$n = 3$	$n = 10$	$n = 100$
Gibbs/k-means	0.18648 (0.10316)	0.09613 (0.09677)	0.02594 (0.04200)
KL	0.03358 (0.04357)	NA	NA
MAP	0.03372 (0.04679)	0.06135 (0.08815)	0.02364 (0.04157)

Table 1: Moyenne (et écart-type) de l'erreur quadratique de la moyenne *a posteriori* sur 100 réplicats et des chaînes de longueur 2000.

Bibliographie

[1] Celeux, G., (1997) Discussion of 'On Bayesian analysis of mixtures models with an unknown number of components' (with discussion), *Journal of Royal Statistical Society: Series B*, 59, 775–776.

[2] Celeux, G., (1998) Bayesian inference for mixtures: the label-switching problem, *R. Payne & P. J. Greens, eds, COMPSTAT 98, Physica, Heidelberg*, 227–232.

[3] Celeux, G., Hurn, M. et Robert, C. P. (2000) Computational and Inferential Difficulties with Mixture Posterior Distributions, *Journal of the American Statistical Association*, 95, 451, 957–970.

[4] Diebolt, J. et Robert, C. P. (1994) Estimation of finite mixture distributions, *Journal of Royal Statistical Society: Series B*, 56, 363–375.

[5] Jasra, A., Holmes, C. C. et Stephens, D. A. (2005) Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling, *Statistical Science*, 20, 1, 50–67.

[6] Sperrin, M. and Jaki, T. and Wit, E. (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models, *Statistics and Computing*, 20, 3, 357–366.

[7] Stephens, M. (1997) Bayesian Methods fo Mixtures of Normal Distribution, D. Phil. thesis, Department of Statistics, University of Oxford.

Inférence bayésienne sur un modèle de mélange a interaction spatiale

Lionel Cucala & Jean-Michel Marin

Institut de Mathématiques et Modélisation de Montpellier, UMR CNRS 5149
Université des Sciences et Techniques du Languedoc
Place Eugène Bataillon
34095 Montpellier cedex 5

Résumé

Nous présentons un algorithme MCMC permettant d'estimer les paramètres d'un modèle de mélange avec interaction spatiale. La difficulté provient de la présence d'une constante de normalisation non calculable : nous nous en affranchissons en utilisant la méthode de Murray *et al.*(2006) consistant à générer des variables aléatoires auxiliaires. Ensuite, nous proposons une technique de sélection du nombre de composantes du mélange basée sur l'approximation de Chib (1995). Nous illustrons ces techniques par l'analyse d'images satellites.

Summary

This article introduces an MCMC algorithm to estimate the parameters of a mixture model with spatial dependence. The main problem comes from an unavailable normalization constant and we skip it thanks to Murray *et al.*(2006)'s method : the idea is to sample from an augmented distribution which involves two auxiliary random variables. Finally, we propose a method to select the number of the components of the mixture model, based on Chib (1995)'s approximation. These techniques have applications in image analysis.

Mots-clés

champs markoviens ; statistique bayésienne ; méthodes de Monte-Carlo par chaînes de Markov ; traitement d'images.

1 Le modèle

Soit $\mathbf{y} = (y_1, \cdots, y_N) \in \{1, \cdots, 256\}^{\otimes N}$ les valeurs des niveaux de gris observés sur les N pixels d'une image. Nous introduisons le champ $\mathbf{z}_k = (z_{1,k}, \cdots, z_{N,k}) \in \mathcal{Z}_k = \{1, \cdots, k\}^{\otimes N}$ donnant la composante de chaque pixel. Soit $\theta_k = (\mu_{1,k}, \cdots, \mu_{k,k}, \sigma_k, \beta_k) \in \mathbb{R}^{\otimes k+2}$ le vecteur des paramètres.

Les valeurs des niveaux de gris sont supposées être des variables aléatoires suivant une loi gaussienne dont les paramètres dépendent de la composante du pixel associé :

$$f(y_i|z_{i,k} = j, \theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\Big[-\frac{1}{2}\Big(\frac{y_i - \mu_{j,k}}{\sigma_k}\Big)^2\Big] .$$

Une hypothèse classique consiste à considérer que les $z_{i,k}$ sont indépendants. Cela correspond au cas classique du modèle de mélange gaussien. Cependant, nous pensons, comme Alston *et al.* (2005), qu'il est nécessaire de prendre en compte les relations de voisinage entre les pixels à travers l'utilisation d'un champ aléatoire markovien pour la variable des composantes. Nous supposerons donc que le champ des composantes $\mathbf{z}_k$ est distribué suivant un modèle de Potts :

$$f(\mathbf{z}_k|\beta_k) = \exp\Big[\beta_k \sum_{l\sim i} \mathbb{1}(z_{i,k} = z_{l,k})\Big]/Z(\beta_k) ,$$

où $\sum_{l\sim i}$ est la somme sur toutes les paires de pixels voisins et

$$Z(\beta_k) = \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \exp\Big[\beta_k \sum_{l\sim i} \mathbb{1}(z_{i,k} = z_{l,k})\Big]$$

est la constante de normalisation. Nous noterons $S(\mathbf{z}_k) = \sum_{l\sim i} \mathbb{1}(z_{i,k} = z_{l,k})$ la statistique exhaustive du modèle de Potts. La distribution jointe est donc

$$f(\mathbf{y}, \mathbf{z}_k|\theta_k) = \prod_{i=1}^{N} f(y_i|z_{i,k}, \theta_k) f(\mathbf{z}_k|\beta_k) ,$$

et celle des observations est

$$f(\mathbf{y}|\theta_k) = \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \prod_{i=1}^{N} f(y_i|z_{i,k}, \theta_k) f(\mathbf{z}_k|\beta_k) .$$

Estimer les paramètres d'un tel modèle, quand on observe uniquement le champ $\mathbf{y}$, est un problème difficile qui peut être résolu dans un cadre bayésien.

2 Echantillonneur de Gibbs

On se place dans le paradigme bayésien et on utilise les lois a priori suivantes :
- $\sigma_k^2 \sim \mathcal{IG}(a, b)$,
- $\mu_k = (\mu_{1,k}, \cdots, \mu_{k,k}) \sim \mathcal{U}(0 \leq \mu_{1,k} \leq \cdots \leq \mu_{k,k} \leq 256)$,
- $\beta_k \sim \mathcal{U}([0, 3])$.

Les valeurs des hyper-paramètres a et b sont choisies de telle que manière que la loi inverse gamma soit très dispersée.

En suivant la méthodologie de Ryden et Titterington (1998), nous construisons un échantillonneur de Gibbs modifié pour générer des réalisations des paramètres et des composantes cachées.

2.1 Etape initiale

Générer σ_k^2, $\mu_{1,k}, \cdots, \mu_{k,k}$ et β_k en utilisant les lois a priori. Puis, générer les $z_{i,k}$ comme si l'on avait à faire à un mélange gaussien, à savoir :

$$z_{i,k} \sim \mathcal{M}(1; \omega_{i,1,k}, \cdots, \omega_{i,k,k}), \quad \forall i = 1, \cdots, N,$$

où

$$\omega_{i,j,k} = \frac{\exp\big[-\frac{1}{2}\big(\frac{y_i - \mu_{j,k}}{\sigma_k}\big)^2\big]}{\sum_{j=1}^{k} \exp\big[-\frac{1}{2}\big(\frac{y_i - \mu_{j,k}}{\sigma_k}\big)^2\big]}.$$

2.2 Etape itérative

Calculer :
- $n_{j,k} = \sum_{i=1}^{N} \mathbb{1}(z_{i,k} = j), \quad \forall j = 1, \cdots, k$;
- $\bar{y}_{j,k} = \frac{1}{n_{j,k}} \sum_{i=1}^{N} y_i \mathbb{1}(z_{i,k} = j), \quad \forall j = 1, \cdots, k$;
- $\hat{s}_{j,k}^2 = \frac{1}{n_{j,k}} \sum_{i=1}^{N} (y_i - \bar{y}_j)^2 \mathbb{1}(z_{i,k} = j), \quad \forall j = 1, \cdots, k$;
- $\hat{s}_k^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_{z_{i,k},k})^2$.

Générer :
- $\sigma_k^2 \sim \mathcal{IG}\big(\frac{N}{2} + a, \frac{\sum_{i=1}^{N} (y_i - \mu_{z_{i,k},k})^2}{2} + b\big)$;
- $\mu_{j,k} \sim \mathcal{N}\big(\bar{y}_{j,k}, \frac{\sigma_k^2}{n_{j,k}}\big) \mathbb{1}(\mu_{j-1,k} \leq \mu_{j,k} \leq \mu_{j+1,k}), \quad \forall j = 1, \cdots, k$;
- β_k est mis à jour par l'algorithme de Murray *et al.* (2006) (voir ci-dessous) ;
- $z_{i,k} \sim \mathcal{M}(1; \omega_{i,1,k}, \cdots, \omega_{i,k,k}), \quad \forall i = 1, \cdots, N$ avec

$$\omega_{i,j,k} = \frac{\exp\big[-\frac{1}{2}\big(\frac{y_i - \mu_{j,k}}{\sigma_k}\big)^2 + \beta_k \sum_{l \sim i} \mathbb{1}(z_{l,k} = j)\big]}{\sum_{j=1}^{k} \exp\big[-\frac{1}{2}\big(\frac{y_i - \mu_{j,k}}{\sigma_k}\big)^2 + \beta_k \sum_{l \sim i} \mathbb{1}(z_{l,k} = j)\big]}.$$

On appelle

$$\left\{\left(\mu_{1,k}^{(t)},\cdots,\mu_{k,k}^{(t)},\sigma_k^{(t)},\beta_k^{(t)},\mathbf{z}_k^{(t)}\right),t=1,\cdots,T\right\}$$

les simulations obtenues après convergence.

2.3 Algorithme de Murray

Cet algorithme permet de simuler le paramètre β_k suivant la distribution conditionnelle $f(\beta_k|z_k)$ même si celle-ci dépend d'une constante de normalisation inconnue. L'idée est de simuler depuis une distribution augmentée qui prend en compte deux variables aléatoires auxiliaires β'_k et z'_k :

$$\pi(\beta_k,\beta'_k,z_k,z'_k)\propto f(\beta_k|z_k)\pi(\beta_k)Q(\beta'_k|\beta_k)f(\beta'_k|z'_k)\,. \tag{1}$$

La distribution a priori de β_k est la loi uniforme sur $[0,3]$ et la distribution $Q(\beta'_k|\beta_k)$ est la loi uniforme sur $[9\beta_k/10;11\beta_k/10]\cap[0,3]$. En fait, l'algorithme d'échange de Murray est une sorte d'algorithme de Metropolis calibré pour (1) et qui fonctionne de la manière suivante (à l'itération t) :

- Mise à jour de β'_k et z'_k : génération de $\beta'_k\sim Q(.|\beta_k^{(t)})$ et $z'_k\sim f(.|\beta'_k)$;
- Echange : $\beta_k^{(t+1)}=\beta'_k$ avec probabilité $\min(1,p^{(t)})$, sinon $\beta_k^{(t+1)}=\beta_k^{(t)}$, avec

$$\begin{aligned}p^{(t)} &= \frac{f(z_k|\beta'_k)\pi(\beta'_k)Q(\beta'_k|\beta_k^{(t)})f(z'_k|\beta_k^{(t)})}{f(z_k|\beta_k^{(t)})\pi(\beta_k^{(t)})Q(\beta_k^{(t)}|\beta'_k)f(z'_k|\beta'_k)}\\ &= \frac{\min(11\beta'_k/10,3)-9\beta'_k/10}{\min(11\beta_k^{(t)}/10,3)-9\beta_k^{(t)}/10}\exp\Big[(\beta_k^{(t)}-\beta'_k)\Big(\sum_{i\sim j}\mathbb{1}(z'_{i,k}=z'_{j,k})-\sum_{i\sim j}\mathbb{1}(z_{i,k}=z_{j,k})\Big)\Big].\end{aligned}$$

Signalons que nous utilisons l'algorithme de Swendsen-Wang pour simuler depuis un modèle de Potts, car le mélange se fait plus rapidement qu'avec un échantillonneur de Gibbs.

3 Sélection du nombre de composantes

Jusqu'à présent, nous avons considéré le nombre de composantes k du modèle connu, ce qui n'est pas le cas en pratique. À la vue des observations, il est donc nécessaire de choisir le nombre adéquat de composantes k. Si l'on considère une distribution a priori uniforme sur l'espace des modèles, le paradigme bayésien consiste à maximiser la vraisemblance intégrée. En effet, si K est la variable alátoire représentant le vrai nombre de composantes, nous pouvons écrire

$$\mathbb{P}(K=k)=\pi(k|\mathbf{y})\propto m_k(\mathbf{y})=\int f(\mathbf{y}|\theta_k)\pi(\theta_k)\mathrm{d}\theta_k\,.$$

La méthode introduite par Chib (1995) pour estimer cette vraisemblance intégrée est basée sur une application du théorème de Bayes. Nous avons :

$$m_k(\mathbf{y}) = \frac{f(\mathbf{y}|\theta_k^*)\pi(\theta_k^*)}{\pi(\theta_k^*|\mathbf{y})}, \tag{2}$$

for all $\theta_k^* \in \mathbb{R}^{\otimes k+2}$.

Une première astuce consiste à utiliser une approximation de Rao-Blackwell basée sur l'échantillonneur de Gibbs pour estimer le denominateur. En effet,

$$\frac{1}{T}\sum_{t=1}^{T} \pi(\theta_k^*|\mathbf{y}, \mathbf{z}_k^{(t)}) \tag{3}$$

est un estimateur consistent de $\pi(\theta_k^*|\mathbf{y})$.

Cependant, alors que l'estimateur (3) est explicite dans un modèle de mélange gaussien, ce n'est pas le cas ici à cause de la dépendance spatiale. En effet, nous avons

$$\pi(\theta_k^*|\mathbf{y}, \mathbf{z}_k) = \pi\big((\mu_{1,k}^*, \cdots, \mu_{k,k}^*, \sigma_k^*)|\mathbf{y}, \mathbf{z}_k\big)\pi(\beta_k^*|\mathbf{z}_k)\,. \tag{4}$$

Le premier terme est explicite et similaire au modéle non-spatial. D'après le théorème de Cochran,

$$\frac{\bar{y}_{j,k} - \mu_{j,k}}{\hat{s}_{j,k}/\sqrt{n_{j,k}}} \sim T_{n_{j,k}-1} \text{ and } N\frac{\hat{s}_k^2}{\sigma_k^2} \sim \chi^2_{N-k},$$

où χ_n^2 et T_n représentent respectivement les distributions du Chi-deux et de Student à n degrés de liberté.

Le second terme de l'expression (4) est donné par

$$\pi(\beta_k^*|\mathbf{z}_k) = \frac{f(\mathbf{z}_k|\beta_k^*)\pi(\beta_k^*)}{\int f(\mathbf{z}_k|\beta_k)\pi(\beta_k)d\beta_k} = \pi\Big(\beta_k^*|S(\mathbf{z}_k)\Big)\,.$$

Pour contourner l'obstacle de la constante de normalisation, nous choisissons d'estimer cette distribution conditionnelle par un estimateur à noyau en nous appuyant sur tous les couples $\Big(\beta_k', S(\mathbf{z}_k')\Big)$ simulés par l'algorithme de Murray. Les largeurs de bande sont sélectionnées suivant les recommandations de Bashtannyk et Hyndman (2001).
Enfin, nous utilisons également l'approximation de Rao-Blackwell pour estimer le numérateur de l'expression (2).
Nous choisirons donc comme nombre de composantes la valeur de k maximisant $m_k(\mathbf{y})$.

Bibliographie

[1] Alston, C., Mengersen, K., Thompson, J., Littlefield, P., Perry, D. et Ball, A. (2005) Extending the Bayesian mixture model to incorporate spatial information in analysing sheep CAT scan images. *Australian Journal of Agricultural Research*, 56, 373–388.

[2] Bashtannyk, D.M., et Hyndman, R.J. (2001) Bandwidth selection for kernel conditional density estimation. *Computational statistics and data analysis*, 36, 279–298.
[3] Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.
[4] Murray, I., Ghahramani, Z. et MacKay, D. (2006) MCMC for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.
[5] Ryden, T. et Titterington, M. (1998) Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, 7, 194–211.

Statistique Mathématique 1

Prévision statistique et inégalité matricielle de type Cramér-Rao, *Emmanuel Onzon*

On commence par un rappel sur le risque quadratique pour l'estimation d'une part et pour la prévision statistique d'autre part. On présente ensuite une inégalité de type Cramér-Rao pour la matrice d'erreur quadratique de prévision. On discute le cas où la borne est atteinte et on présente un prédicteur sans biais efficace pour le processus de Poisson bivarié.

On the estimation of a restricted location parameter for symmetric distributions, *Ouassou Idir*

On considère le problème de l'estimation de la moyenne d'une distribution à symétrie sphérique unidimensionnelle quand celle-ci est restreinte à un intervalle [a ; b]. Sans perte de généralité, nous supposons que a =-m, b = m, avec m $>$ 0. Les questions intéressantes concernent la performance du fréquentiste d'estimateur Bayesien, telles que la détermination de l'estimateur Bayesien qui améliore l'estimateur de maximum de vraisemblance (mle).

Exact nonparametric two-sample homogeneity tests for possibly discrete distributions, *Abdeljelil Farhat and Jean-Marie Dufour*

In this paper, we study several tests for the equality of two unknown distributions. Two are based on empirical distribution functions, three others on nonparametric probability density estimates, and the last ones on differences between sample moments. We suggest controlling the size of such tests (under nonparametric assumptions) by using permutational versions of the tests jointly with the method of Monte Carlo tests properly adjusted to deal with discrete distributions. We also propose a combined test procedure, whose level is again perfectly controlled through the Monte Carlo test technique and has better power properties than the individual tests which are combined. Finally, in a simulation experiment, we show that the technique suggested provides perfect control of test size and that the new tests proposed can yield sizeable power improvements.

Sur la convergence asymptotique des M-estimateurs pondérés, *Mohammed El Asri*

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ issue de la même loi paramétrique P_θ, avec θ dans un ouvert Θ de $\mathbb{R}^d$. Soit $\rho(x, \theta)$ une fonction mesurable en x, et telle que $\theta_0 := \underset{\theta\in\Theta}{\operatorname{argmin}} \int \rho(x, \theta) dP_\theta$ existe et soit unique. Nous étudions dans ce travail la classe des M-estimateurs pondérés minimisant la fonction objective $\frac{1}{n}\sum_{i=1}^{n} \omega_i \rho(X_i, \theta)$.
La médiane spatiale est un M-estimateur qui présente de très bonnes qualités de robustesse (avec un point de rupture maximal) et de bonnes propriétés de convergence asymptotique [Brown (1983), Chakraborty et Chaudhuri (1996)].

Nevalainen et al. (2006) et (2007) ont étudié, dans le cas des variables clusterisées, deux types de médianes spatiales : pondérée et non-pondérée. Ils montrent la consistance, la normalité asymptotique de leurs estimateurs et le fait que la pondération permette d'améliorer l'efficacité relative de la médiane spatiale. Le résultat reste valable même dans le cas i.i.d où la corrélation intracluster est nulle.
Il paraît pertinent, de généraliser ces résultats pour la classe des M-estimateurs pondérés. Dans un premier temps, nous explicitons les hypothèses qui permettent d'étudier le comportement asymptotique des M-estimateurs pondérés (consistance et normalité). Enfin, nous détaillons deux exemples d'application : la médiane spatiale et l'estimateur de Huber pondérés.

Vraisemblance empirique pour les séries temporelles périodiques, *Hugo Harari-Kermadec and Jacey Leskow*

Dans cette présentation, nous proposons d'utiliser la vraisemblance empirique pour des données dépendantes suivant un modèle périodique. L'originalité de cette étude tient à la non stationnarité des séries étudiées. Pour simplifier la présentation, nous étudierons le cas d'un processus auto-régressif périodique d'ordre 1, un P-AR(1). La construction de blocs de données respectant la structure de dépendance est centrale ici : les données étant P périodique, nous proposons de choisir des blocs de longueur proportionnelle à P.

Prévision statistique et inégalité matricielle de type Cramér-Rao

Emmanuel Onzon - UPMC
emmanuel.onzon@upmc.fr

Laboratoire de Statistique Théorique et Appliquée (LSTA)
Université Pierre et Marie Curie – Paris 6
4, place Jussieu, 75252 Paris cedex 05

RESUME

On commence par un rappel sur le risque quadratique pour l'estimation d'une part et pour la prévision statistique d'autre part. On présente ensuite une inégalité de type Cramér-Rao pour la matrice d'erreur quadratique de prévision. On discute le cas où la borne est atteinte et on présente un prédicteur sans biais efficace pour le processus de Poisson bivarié.

Mots-clés

Inégalité de Cramér-Rao, Prédicteur efficace, Prévision, Information de Fisher

ABSTRACT

We begin firstly with a reminder about the quadratic risk for estimation and secondly for statistical prediction. Then we present a Cramér-Rao type inequality for the quadratic prediction error matrix. We discuss the attainment of the bound and we present an unbiased efficient predictor for the bivariate Poisson Process.

Keywords

Cramér-Rao inequality, Efficient predictor, Prediction, Fisher information

1 Erreur quadratique de prévision

On considère la théorie statistique de la prévision comme une extension de la théorie statistique de l'estimation telle que présentée par Bosq et Blanke (2007). On rappelle brièvement le cadre de l'estimation avant d'aborder son extension à celui de la prévision. Dans la théorie de l'estimation, le statisticien observe la réalisation d'une variable aléatoire X dont la distribution est inconnue mais supposée appartenir à une famille de mesures de probabilité $(\mathbb{P}_\theta)_{\theta\in\Theta}$, où $\theta \in \Theta$ est le paramètre inconnu de la famille. Dans ce cadre, le problème est d'estimer le paramètre θ ou une fonction $g(\theta)$ de ce paramètre. Pour cela le statisticien construit un *estimateur* qui est une fonction de l'observation X. Un moyen

commode de mesurer la performance d'un estimateur $T(X)$ donné, qui admet un moment d'ordre 2, est d'utiliser l'erreur quadratique

$$\mathbb{E}_\theta\big(T(X) - g(\theta)\big)^2.$$

Dans le cas où g est à valeur multidimensionnelle on peut utiliser la matrice d'erreur quadratique

$$\mathbb{E}_\theta\big(T(X) - g(\theta)\big)\big(T(X) - g(\theta)\big)^{\mathrm{T}}.$$

Pour comparer deux matrices d'erreur A et B on peut utiliser l'ordre partiel suivant. $A \leqslant B$ ssi $B - A$ est semi-definie positive.

Dans le cadre de la prévision on considère un processus aléatoire $(Z_t)_{t\in I}$, avec I l'ensemble des dates qui peut être discret ou continu. On suppose que sa loi appartient à la famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$ et qu'on observe le processus sur un sous-ensemble $J \subset I$. La variable observée est donc $X = (Z_t)_{t\in J}$. Si on suppose par exemple qu'on connait le processus jusqu'au temps T on a donc $X = (Z_t)_{t\leqslant T}$, si on suppose qu'on observe le processus uniquement à la date T alors on a $X = Z_T$. On cherche à prédire une valeur du processus à une date ultérieure aux observations, par exemple Z_{T+h}, au moyen d'un prédicteur $p(X)$ construit avec les données observées X. L'erreur quadratique de prévision est alors

$$\mathbb{E}_\theta\big(p(X) - Z_{T+h}\big)^2.$$

On a la décomposition suivante de l'erreur quadratique

$$\mathbb{E}_\theta\big(p(X) - Z_{T+h}\big)^2 = \mathbb{E}_\theta\big(p(X) - \mathbb{E}_\theta[Z_{T+h}|X]\big)^2 + \mathbb{E}_\theta\big(\mathbb{E}_\theta[Z_{T+h}|X] - Z_{T+h}\big)^2.$$

On remarque alors que minimiser l'erreur par rapport à Z_{t+h} revient à minimiser l'erreur par rapport à $\mathbb{E}_\theta[Z_{T+h}|X]$ qui est une variable aléatoire X-mesurable qui est fonction de θ. Ceci justifie d'étudier l'erreur de $p(X)$ par rapport à une quantité quelconque $g(X,\theta)$

$$\mathbb{E}_\theta\big(p(X) - g(X,\theta)\big)^2,$$

où $x \mapsto g(x,\theta)$ est une fonction mesurable pour tout $\theta \in \Theta$.

Lorsque X est multidimensionnel l'erreur correspondante est la matrice

$$\mathbb{E}_\theta\big(p(X) - g(X,\theta)\big)\big(p(X) - g(X,\theta)\big)^{\mathrm{T}}.$$

Enfin on définit la notion de biais pour la prévision de manière analogue à l'estimation. Un prédicteur $p(X)$ est dit sans biais pour prédire Z_{T+h} si $\mathbb{E}_\theta p(X) = \mathbb{E}_\theta Z_{T+h}$ pour tout $\theta \in \Theta$ et il est sans biais pour prédire $g(X,\theta)$ si $\mathbb{E}_\theta p(X) = \mathbb{E}_\theta g(X,\theta)$ pour tout $\theta \in \Theta$.

2 Inégalité de Cramér-Rao pour la prévision

Soit θ le paramètre d'une famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$ et X une variable aléatoire observée de loi $\mathbb{P}_\theta$. On considère le problème d'estimation de $g(\theta)$, une fonction différentiable du paramètre θ, par un estimateur sans biais $T(X)$. L'inégalité de Cramér-Rao pour l'estimation donne une borne inférieure pour l'erreur quadratique de $T(X)$ lorsque la famille admet une densité $f(x,\theta)$ qui satisfait certaines conditions de régularité que nous ne rappelons pas ici mais qu'on trouve dans les ouvrages de référence sur l'estimation ponctuelle tels que Lehmann et Casella (1998). On rappelle l'inégalité dans le cas où l'estimateur sans biais $T(X)$ estime une quantité $g(\theta)$ avec g une fonction différentiable et telle que θ et $g(\theta)$ sont multidimensionnels (inégalité matricielle avec $A \leqslant B$ ssi $B-A$ est semi-définie positive).

$$\mathbb{E}_\theta\big(T(X)-g(\theta)\big)\big(T(X)-g(\theta)\big)^{\mathrm{T}} \geqslant G(\theta)I(\theta)^{-1}G(\theta)^{\mathrm{T}},$$

où $G(\theta)$ est la matrice jacobienne de g et $I(\theta)$ est la matrice d'information de Fisher

$$I(\theta) = \mathbb{E}_\theta\big(\nabla_\theta \ln f(X,\theta)\big)\big(\nabla_\theta \ln f(X,\theta)\big)^{\mathrm{T}},$$

où on a noté ∇_θ l'opérateur gradient par rapport à la variable multidimensionnelle θ.

Yatracos (1992) a donné une généralisation de cette inégalité dans le cadre de la prévision statistique pour un paramètre et une variable à prédire unidimensionnels. Nayak (2002) généralise au cas multidimensionnel en donnant une inégalité matricielle. On introduit des hypothèses avant d'énoncer cette inégalité.

Hypothèses 1 $\Theta \subset \mathbb{R}^d$ est un ouvert, le modèle associé à X est dominé par une mesure σ-finie μ, X admet une densité $f(x,\theta)$ par rapport à μ telle que $\{x \ : \ f(x,\theta) > 0\}$ ne dépend pas de θ et $\nabla_\theta f(x,\theta)$ existe. Finalement l'information de Fisher relative à X

$$I(\theta) = \mathbb{E}_\theta[\dot{L}_\theta \dot{L}_\theta^{\mathrm{T}}] \quad \text{avec} \quad \dot{L}_\theta = \nabla_\theta \ln f(X,\theta),$$

satisfait $\det\big(I(\theta)\big) \neq 0$ et tous ses coefficients sont finis.

On note J_θ l'opérateur matrice jacobienne par rapport à la variable multidimensionnelle θ.

Théorème 1 *Soient $p : \mathcal{X} \to \mathbb{R}^k$ un prédicteur sans biais de $g(X,\theta)$ et $g : \mathcal{X}\times\Theta \to \mathbb{R}^k$ telle que pour tout $\theta \in \Theta$ la fonction $x \mapsto g(x,\theta)$ est mesurable, $g(X,\theta) \in \mathrm{L}^2(\mathbb{P}_\theta)$, $\Theta \subset \mathbb{R}^d$ et pour tout x la fonction $\theta \mapsto g(x,\theta)$ est différentiable sur Θ. Si les hypothèses 1 sont satisfaites et si en tout point θ de Θ les égalités*

$$\int g(x,\theta) f(x,\theta)\, d\mu(x) = \mathbb{E}_\theta\big[g(X,\theta)\big], \tag{1}$$

$$\int p(x) f(x,\theta)\, d\mu(x) = \mathbb{E}_\theta\big[g(X,\theta)\big], \tag{2}$$

sont dérivables sous le signe intégral par rapport à chacune des composantes de θ*, alors*

$$\mathbb{E}_\theta\big(p(X) - g(X,\theta)\big)\big(p(X) - g(X,\theta)\big)^{\mathrm{T}} \;\geqslant\; G(\theta)\, I(\theta)^{-1}\, G(\theta)^{\mathrm{T}}, \tag{3}$$

où on a noté $G(\theta) = \mathbb{E}_\theta\big(\mathrm{J}_\theta g(X,\theta)\big)$.

3 Prédicteur efficace

Dans cette section on étudie le cas où la borne de l'inégalité (3) est atteinte. Les résultats suivants apparaissent dans Onzon (2011).

Proposition 2 *On suppose les hypothèses du théorème 1. L'inégalité (3) devient une égalité si et seulement si*

$$p(X) = g(X,\theta) + G(\theta) I(\theta)^{-1} \dot{L}_\theta, \quad \mathbb{P}_\theta\text{-p.s.}$$

On rappelle qu'un estimateur sans biais dont l'erreur quadratique atteint la borne de Cramér-Rao est dit *efficace*. Suivant cette convention on appelle *prédicteur efficace* sans biais un prédicteur sans biais dont l'erreur quadratique atteint la borne de l'inégalité (3). Lorsque le prédicteur est efficace et que $k = d$ la densité satisfait certaines conditions données dans le théorème suivant.

Théorème 3 *On suppose que les hypothèses 1 sont satisfaites et l'égalité (1) est dérivable sous le signe intégral. Soit* $p : \mathcal{X} \to \mathbb{R}^k$ *un prédicteur* efficace *sans biais de* $g(X,\theta)$ *et* $g : \mathcal{X} \times \Theta \to \mathbb{R}^k$ *une fonction telle que pour tout* $\theta \in \Theta$ *la fonction* $x \mapsto g(x,\theta)$ *est mesurable et* $g(X,\theta) \in \mathrm{L}^2(P_\theta)$*,* $\Theta \subset \mathbb{R}^d$ *et pour tout* x *la fonction* $\theta \mapsto g(x,\theta)$ *est différentiable sur* Θ*. On note* $G(\theta) = \mathbb{E}_\theta\big(\mathrm{J}_\theta g(X,\theta)\big)$.

Si pour tout θ *la matrice* $G(\theta)$ *est inversible et qu'il existe une fonction différentiable* $A : \Theta \to \mathbb{R}^k$*, telle que* $\big(\mathrm{J}_\theta A(\theta)\big)^{\mathrm{T}} = I(\theta)G(\theta)^{-1}$*, alors il existe* $B : \mathcal{X} \times \Theta \to \mathbb{R}$ *une fonction différentiable par rapport à* $\theta \in \Theta$ *telle que*

$$f(x,\theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x,\theta)\big), \quad \forall \theta \in \Theta \text{ et } \mathbb{P}_{X,\theta}\text{-p.t. } x \in \mathcal{X},$$

et $\nabla_\theta B(x,\theta) = \big(\mathrm{J}_\theta A(\theta)\big)^{\mathrm{T}} g(x,\theta)$*,* $\forall \theta \in \Theta$ *et* $\mathbb{P}_{X,\theta}$*-p.t.* $x \in \mathcal{X}$.

Remarque La famille de densités $f_\theta(x) = \exp\{\langle A(\theta), p(x)\rangle - B(x,\theta)\}$ n'est pas une famille exponentielle car $B(x,\theta)$ peut ne pas être une somme $B_1(x) + B_2(\theta)$. Néanmoins lorsque g dépend seulement de θ, l'égalité $\nabla_\theta B(x,\theta) = \big(\mathrm{J}_\theta A(\theta)\big)^{\mathrm{T}} g(\theta)$ implique qu'il existe B_1 et B_2 telles que $B(x,\theta) = B_1(x) + B_2(\theta)$, pour tout $\theta \in \Theta$ et $x \in \mathcal{X}$. Ce cas est celui où le cadre de la prévision se réduit à celui de l'estimation.

Le théorème suivant est une réciproque

Théorème 4 *On suppose que les hypothèses 1 sont satisfaites. Soit $g : \mathcal{X} \times \Theta \to \mathbb{R}^k$ une fonction telle que pour tout $\theta \in \Theta$ la fonction $x \mapsto g(x, \theta)$ est mesurable et $g(X, \theta) \in \mathrm{L}^2(P_\theta)$ et $\Theta \subset \mathbb{R}^d$. On suppose que X admet la densité*

$$f(x, \theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x, \theta)\big), \quad \theta \in \Theta, \tag{4}$$

avec $p : \mathcal{X} \to \mathbb{R}^k$ une fonction mesurable et $A : \Theta \to \mathbb{R}^d$ deux fois différentiable de matrice jacobienne inversible et $B : \mathcal{X} \times \Theta \to \mathbb{R}$ deux fois différentiable par rapport à $\theta \in \Theta$, où A et B satisfont $g(X, \theta) = \big(\mathrm{J}_\theta A(\theta)\big)^{-1\mathrm{T}} \nabla_\theta B(X, \theta)$.

Si $\int f(x, \theta)\, d\mu(x)$ est deux fois dérivable sous le signe intégral, alors $p(X)$ est un prédicteur efficace sans biais de $g(X, \theta)$.

Remarque Toute densité $f(x, \theta)$ qui ne s'annule pas peut s'écrire comme (4), en choisissant $A(\theta) = \theta$, $p(x) = 0$ et $B(x, \theta) = -\log(f(x, \theta))$. Mais sous cette forme, la quantité à prédire donnée par le théorème n'est pas nécessairement intéressante. Le théorème est utile quand il est possible d'écrire la densité de telle sorte que la quantité $\big(\mathrm{J}_\theta A(\theta)\big)^{-1\mathrm{T}} \nabla_\theta B(X, \theta)$ est intéressante à prédire.

4 Prévision d'un processus de Poisson bivarié

On considère le processus de Poisson bivarié $\big(N_t\big)_{t \geqslant 0}$ en suivant la définition de Marshall et Olkin (1967) et on note $\forall t \geqslant 0,\ N_t = \begin{pmatrix} N_{1,\,t} \\ N_{2,\,t} \end{pmatrix}$. Il est markovien et ses accroissements sont indépendants et stationnaires. Le paramètre du modèle est $\theta = (\lambda_1,\, \lambda_2,\, \lambda_3) \in {\mathbb{R}_+^*}^3$ avec $\lambda_3 \leqslant \min(\lambda_1,\, \lambda_2)$. La loi est

$$f(x, \theta) = \mathbb{P}_\theta\Big(N_t = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\Big) = e^{-(\lambda_1+\lambda_2-\lambda_3)t} \sum_{k=0}^{\min(x_1,\, x_2)} \frac{\lambda_3^k (\lambda_1 - \lambda_3)^{x_1-k} (\lambda_2 - \lambda_3)^{x_2-k} t^{x_1+x_2-k}}{k!(x_1 - k)!(x_2 - k)!},$$

avec $x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^{\mathrm{T}} \in \mathbb{N}^2$.

Dans ce qui suit on étudie le problème de prédire N_{T+h} en supposant qu'on connait le processus uniquement à la date T (donc la variable observée est $X = N_T$). D'après Kocherlakota et Kocherlakota (1992) l'inverse de la matrice d'information relative à l'observation X_T est

$$I(\theta)^{-1} = \frac{1}{T} \begin{pmatrix} \lambda_1 & \lambda_3 & \lambda_3 \\ \lambda_3 & \lambda_2 & \lambda_3 \\ \lambda_3 & \lambda_3 & \delta \end{pmatrix}.$$

(avec δ donné dans Kocherlakota et Kocherlakota (1992))

Pour tout $t \geqslant 0$, $N_{1,\,t}$ et $N_{2,\,t}$ sont des variables de Poisson de paramètres respectifs $\lambda_1 t$ et $\lambda_2 t$, et de covariance égale à $\lambda_3 t$.

L'espérance conditionnelle de N_{T+h} sachant N_T est

$$\mathbb{E}_\theta^{N_T} N_{T+h} = \mathbb{E}_\theta^{N_T}[N_{T+h} - N_T + N_T] = \mathbb{E}_\theta[N_{T+h} - N_T] + N_T = \mathbb{E}_\theta N_h + N_T = h \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + N_T.$$

Ceci et le théorème de Lebesgue prouvent la dérivabilité de (1) sous le signe intégral.

λ_1 et λ_2 s'estiment sans biais par $\widehat{\lambda}_1 = \frac{N_{1,T}}{T}$ et $\widehat{\lambda}_2 = \frac{N_{2,T}}{T}$ respectivement. Ce qui incite à poser le prédicteur $p(N_T) = h \begin{pmatrix} \widehat{\lambda}_1 \\ \widehat{\lambda}_2 \end{pmatrix} + N_T = \frac{h}{T} N_T + N_T = \frac{T+h}{T} N_T$ qui est sans biais. L'égalité (2) est dérivable sous le signe intégral par rapport à λ_1, λ_2 et λ_3, en utilisant le théoreme de Lebesgue.

Bosq et Blanke (2007) ont montré que $p(N_T)$ est un prédicteur efficace dans le cas univarié. Nous allons voir que c'est aussi vrai dans le cas bivarié. L'erreur quadratique du prédicteur par rapport à l'espérance conditionnelle est

$$\begin{aligned} R_T &= \mathbb{E}_\theta\big(p(N_T) - \mathbb{E}_\theta^{N_T} N_{T+h}\big)\big(p(N_T) - \mathbb{E}_\theta^{N_T} N_{T+h}\big)^{\mathrm{T}} \\ &= h^2\, \mathbb{E}_\theta\Big(\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} - \frac{N_T}{T}\Big)\Big(\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} - \frac{N_T}{T}\Big)^{\mathrm{T}} = \frac{h^2}{T} \begin{pmatrix} \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_2 \end{pmatrix}. \end{aligned}$$

On calcule maintenant la borne de Cramér-Rao $G(\theta) I(\theta)^{-1} G(\theta)^{\mathrm{T}}$.

$$G(\theta) = \mathbb{E}_\theta\big(\mathrm{J}_\theta \mathbb{E}_\theta^{N_T} N_{T+h}\big) = \begin{pmatrix} h & 0 & 0 \\ 0 & h & 0 \end{pmatrix},$$

$$G(\theta) I(\theta)^{-1} G(\theta)^{\mathrm{T}} = \frac{h^2}{T} \begin{pmatrix} \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_2 \end{pmatrix} = R_T.$$

La borne de Cramér-Rao est atteinte, $p(N_T)$ est un prédicteur efficace.

Bibliographie

[1] Bosq, D. et Blanke, D., (2007) *Inference and prediction in large dimensions*, Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester.

[2] Kocherlakota, S., Kocherlakota, K., (1992) *Bivariate discrete distributions*, Vol. 132 of Statistics : Textbooks and Monographs. Marcel Dekker Inc., New York.

[3] Lehmann, E. L., Casella, G., (1998), *Theory of point estimation, 2nd Edition*, Springer Texts in Statistics. Springer-Verlag, New-York.

[4] Marshall, A. W., Olkin, I., (1967), A generalized bivariate exponential distribution. *Journal of Applied Probability*, 4, 291–302.

[5] Nayak, T. K., (2002), Rao-Cramer type inequalities for mean squared error of prediction. *American Statistician*, 56 (2), 102–106.

[6] Onzon, E., (2011), Multivariate Cramér-Rao inequality for prediction and efficient predictors. *Statistics and Probability letters*, 81, 429–437.

[7] Yatracos, Y. G., (1992), On prediction and mean squared error. *The Canadian Journal of Statistics*, 20 (2), 187–200.

On the estimation of a restricted location parameter for symmetric distributions

Idir Ouassou, En colaboration avec Eric Marchand, Amir T. Payandeh & François Perron.

Ecole Nationale des Sciences Appliques
Avenue Abdelkrim El Khattabi
BP 575, Marrakech, MAROC
idir.ouassou@ensa.ac.ma.

Résumé

On considère le problème de l'estimation de la moyenne d'une distribution à symétrie sphérique unidimensionnelle quand celle-ci est restreinte à un intervalle $[a, b]$. Sans perte de généralité, nous supposons que $a = -m$, $b = m$, avec $m > 0$.

Consider the problem of estimating under squared-error loss, based on an observable X, the median θ of a spherically symmetric univariate model where θ is known to be restricted to an interval $[a, b]$. Without loss of generality, we set $a = -m$, $b = m$, with $m > 0$. Interesting questions pertain to the frequentist performance of Bayesian estimators, such as the determination or description of Bayesian estimators that improve upon the benchmark (but inadmissible; e.g., Charras and van Eeden, 1993) maximum likelihood estimator (δ_{mle}). With respect to the last question, findings have been obtained, for normal models and among others, by Casella and Strawderman (1981); Gatsonis, MacGibbon and Strawderman (1987); and Marchand and Perron (2001); and for more general models by Moors (1981, 1985) and Marchand and Perron (2005, 2007).

The results of Marchand and Perron (2001), which are developed for multivariate versions with $X \sim N_p(\theta, \sigma^2 I_p)$, $\|\theta\| \leq m$, and known σ^2, will be of particular relevance to us. Namely, their results for $p = 1$ imply that:

(i) all Bayes estimators with respect to a symmetric about 0 prior dominate δ_{mle} as soon as $m \leq c_0\sigma$ with $c_0 \approx 0,4837$;

(ii) the Bayes estimator δ_{BU} with respect to a boundary uniform prior on $\{-m, m\}$ dominates δ_{mle} as soon as $m \leq \sigma$;

(iii) the Bayes estimator δ_U with respect to a fully uniform on $[-m, m]$ prior dominates δ_{mle} as soon as $m \leq c_U\sigma$ with $c_U \approx 0,5230$;

(iv) a Bayes estimator δ_π with respect to an absolutely continuous, symmetric and logconcave density π on $[-m, m]$, with $\pi(\cdot)$ also nondecreasing on $[0, m]$, dominates δ_{mle} for $m \leq c_\pi \sigma$ with $c_\pi \geq c_U$ (see their Remark 4, part 1)

Extensions of **(ii)** to a large class of symmetric models were also obtained by Marchand and Perron (2005, 2007).
The main original contributions in this paper are explicit extensions of **(i)**, **(iii)**, and **(iv)** to various other univariate, unimodal, and symmetric models. Although the findings of Marchand and Perron (2001) do suggest that the existence of such dominance results is perhaps not surprising, it remains particularly interesting that a unified development for logconcave densities with a first derivative which is logconvex on $(0, 2m)$, such as the one given here, is possible and leads to reasonable simple dominance conditions. Moreover, analogous dominance results to **(i)** are established, outside this framework; first to Student distributions providing a stronger general Bayesian result than the one obtained by Marchand and Perron (2005) which is limited to the estimator δ_{BU} and **(ii)**; secondly to logconvex on (θ, ∞) densities which will include scale mixtures of Laplace densities. Various other illustrations are given. These include : examples of dominating Bayesian estimators, numerical evaluations of the cutoffs points (e.g., c_π as in $m \leq c_\pi$) for dominance, applications to truncated linear and truncated linear minimax estimators and a multivariate Student dominance result.
As in Marchand and Perron (2001), the technical arguments relied upon to arrive at these findings begin with conditional risk decompositions of the type $|X| \leq m$, $|X| > m$, and $|X| = r$, $r > 0$. Perron (2003) exploited these types of decompositions to obtain analogous dominance results for estimating a Binomial proportion p when $|p - \frac{1}{2}|$ is constrained above, while Marchand and Perron (2005, 2007) made use of the conditioning on $|X| = r$ to obtain their previous extensions of **(ii)**.
Along with the recent reviews of estimation problems in restricted parameter spaces given by Marchand and Strawderman (2004) or van Eeden (2006), a renewed interest in restricted parameter space inference has occurred with problems arising in particle physics as reported upon by Mandelkern (2002). An additional motivation for the type of problem considered here comes from a connection with inference problems in the presence of additional information, such as the problem of estimating a mean θ_1 based on observables Y_1, Y_2 with $E(Y_i) = \theta_i; i = 1, 2$, subject to the additional information that $|\theta_1 - \theta_2| \leq c$ for some known constant c (e.g., Marchand and Strawderman, 2004; van Eeden and Zidek, 2004).

Keywords and phrases : Maximum likelihood estimator, restricted parameter space, Bayes estimator, squared error loss, dominance, symmetric location families, Cauchy and Student models, logconcave densities, logconvex densities, scale mixture of Laplace densities.

References

[1] Berger, J. O. Minimax estimation of a multivariate normal mean under polynomial loss. *Journal of Multivariate Analysis*, **8**: 173–180, 1978.

[2] Casella, G. & Strawderman, W.E. Estimating a bounded normal mean. *Annals of Statistics*, **9**: 870-878, 1981.

[3] Charras, A. & van Eeden, C.. Bayes and admissibility properties of estimators in truncated parameter spaces, *Canadian Journal of Statistics*, **19**: 121-134, 1981.

[4] Dou, Y. & van Eeden, C. Comparisons of the performances of estimators of a bounded normal mean under squared-error loss. *Technical Report 223, Department of Statistics, The University of British Columbia* 2006.

[5] Gatsonis, C, MacGibbon, B. & Strawderman, W.E.. On the estimation of a restricted normal mean, *Statistics and Probability Letters*, **6**: 21-30, 1987

[6] Idir Ouassou, Eric Marchand, Amir T. Payandeh & François Perron . On the estimation of a restricted location parameter for symmetric distributions, *Journal of the Japan Statistical Society*, **Vol. 38** : 1-17, 2008.

[7] Kubokawa, T. . Estimation of bounded location and scale parameters. *Journal of the Japanese Statistical Society*, **35**: 221-249, 2005.

[8] Mandelkern, M. Setting Confidence Intervals for Bounded Parameters with discussion, *Statistical Science*, **17**: 149-172,2002.

[9] Marchand, É. & Perron, F. Improving on the MLE of a bounded normal mean. *Annals of Statistics*, **29**: 1078-1093,2001.

[10] Marchand, É. & Perron, F. Improving on the MLE of a bounded mean for spherical distributions. *Journal of Multivariate Analysis*, **92**: 227-238,2005.

[11] Marchand, É. & Perron, F. Estimating a bounded parameter for symmetric distributions. *Annals of the Institute of Mathematical Statistics*, to appear 2007.

[12] Marchand, É. & Strawderman, W. E. Estimation in restricted parameter spaces: A review. *Festschrift for Herman Rubin*, IMS Lecture Notes-Monograph Series, **45**: pp. 21-44, 2004.

[13] Moors, J.J.A. Inadmissibility of linearly invariant estimators in the truncated parameter spaces, *Journal of the American Statistical Association*, **76**: 910-915, 1981.

[14] Moors, J.J.A. Estimation in truncated parameter spaces. *Ph.D. thesis, Tilburg University* 1985.

[15] Perron, F. Improving on the mle of p for a Binomial(n, p) when p is around $1/2$. *Festschrift for Constance van Eeden*, IMS Lecture Notes-Monograph Series, **42**: pp. 45-64, 2003.

[16] van Eeden, C. Restricted parameter space problems. Admissibility and minimaxity properties. *Lecture Notes in Statistics,* **188**, Springer 2006.

[17] van Eeden, C. & Zidek, J.V. Combining the data from two normal populations to estimate the mean of one when their difference is bounded. *Journal of Multivariate Analysis*, **88**: 19-46, 2004.

Exact nonparametric two-sample homogeneity tests for possibly discrete distributions

Jean-Marie Dufour* and Abdeljelil Farhat**

** Department of Economics, McGill University, Montréal, Canada,*
*** Unité de recherche EAS-Mahdia Faculté des sciences économiques et de gestion de Mahdia, Université de Monastir, Tunisie,*

Abstract

In this paper, we study several tests for the equality of two unknown distributions. Two are based on empirical distribution functions, three others on nonparametric probability density estimates, and the last ones on differences between sample moments. We suggest controlling the size of such tests (under nonparametric assumptions) by using permutational versions of the tests jointly with the method of Monte Carlo tests properly adjusted to deal with discrete distributions. We also propose a combined test procedure, whose level is again perfectly controlled through the Monte Carlo test technique and has better power properties than the individual tests which are combined. Finally, in a simulation experiment, we show that the technique suggested provides perfect control of test size and that the new tests proposed can yield sizeable power improvements.

Key words: nonparametric methods; two-sample problem; discrete distribution; discontinuous distribution; goodness-of-fit test; Kolmogorov-Smirnov test; Cramér-von Mises; kernel density estimator; exact test; permutation test; Monte Carlo test; bootstrap; combined test procedure; induced test.

Résumé

Dans ce texte, nous étudions plusieurs tests pour l'egalité de deux distributions inconnues. Deux de ces tests sont basés sur des fonctions de distribution empiriques, trois autres sur des estimateurs non-paramétriques de fonctions de densité, et les trois derniers sur des moments empiriques. Nous proposons de contrôler la taille des tests (sous des hypothèses non-paramétriques) en employant des versions permutationnelles de ces tests conjointement avec la méthode des tests de Monte Carlo ajustée pour tenir compte de la possibilité de distributions discontinues. Nous proposons aussi une méthode pour combiner plusieurs de ces tests, le niveau de ces procédures étant aussi contrôlé par la technique des tests de Monte Carlo, laquelle possède de meilleures propriétés de puissance que les tests individuels combinés. Finalement, nous montrons dans une étude de simulation que la technique suggérée contrôle parfaitement la taille des différents tests considérés et que les nouveaux tests proposés peuvent fournir de notables améliorations de puissance.

Mots clés: méthodes non-paramétriques; problème des deux échantillons; distribution discrète; distribution discontinue; test d'ajustement; test de Kolmogorov-Smirnov; estimateur à noyau pour une densité; test exact; test de permutations; test de Monte Carlo; bootstrap; test combiné; test induit.

Introduction and summary

An important problem in statistics consists in testing whether the distributions of two random variables are identical against the alternative that they differ in some way. Specifically, consider two random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ such that $F(x) = \mathsf{P}[X_i \leq x]$, $i = 1, \ldots, n$, and $G(x) = \mathsf{P}[Y_j \leq x]$, $j = 1, \ldots, m$. We shall not impose here additional restrictions on the form of the cumulative distribution functions (cdf) F and G, which may be continuous or discrete. The problem consists in testing the null hypothesis

$$H_0 : F = G \tag{1}$$

against the alternative

$$H_1 : F \neq G\,. \tag{2}$$

H_0 is a nonparametric hypothesis, so testing H_0 requires a distribution-free procedure. Thus, many users who have to make such a confrontation resort to a goodness-of-fit test, usually the two-sample Kolmogorov-Smirnov (KS) test [Smirnov(1948)] or the Cramér-von Mises (CM) test [Lehmann(1951), Rosenblatt(1952) and Fisz(1960)]. Other procedures that have been suggested include permutation tests based on L_1 and L_2 distances between kernel-type estimators of the relevant probability density functions (pdf)

[Allen(1997)] and tests based on the difference of the means of the two samples considered [Pitman(1937), Dwass(1957), Efron-Tibshirani(1993)]. Except for the last procedure, which is designed to have power against samples that differ through their means, the exact and limiting distributions of the test statistics are not standard, and tables for the exact distributions are only available for a limited number of sample sizes. Thus these tests are usually performed with the help of tables based on asymptotic distributions. This leads to procedures that do not have the targeted size (which can easily be too small or too large) and may have low power.

In this paper, we aim at finding test procedures with two basic features. Namely, the latter should be: (1) truly distribution-free, irrespective of whether the underlying distribution F is discrete or continuous, and (2) exact in finite samples (i.e., they must achieve the desired size even for small samples). In this respect, it is important to note that the finite and large sample distributions of usual test statistics are not necessarily distribution-free under H_0. In particular, while the KS and CM statistics are distribution-free when the observations are independent and identically distributed ($i.i.d.$) with a continuous distribution, this is not anymore the case when they follow a discrete distribution. For the statistics based on kernel-type density estimators, distribution-freeness does not obtain even for $i.i.d$ observations with a continuous distribution. This difficulty can be relaxed by considering a permutational version of these tests which uses the fact that all permutations of the pooled observations are equally likely when the observations are $i.i.d$ with a continuous distributions. The latter property, however, does not hold when the observations follow a discrete distribution. So none of the procedures proposed to date for testing H_0 satisfies the double requirement of yielding a test that is both distribution-free and exact. For further discussion of permutation tests, the reader may consult Dufour-Hallin(1993) and Pesarin(2001).

Given recent progress in computing power, a way to solve this difficulty consists in using simulation-based methods, such as bootstrapping or Monte Carlo tests. The bootstrap technique however does not ensure that the level will be fully controlled in finite samples [for further discussion of bootstrapping, see Efron-Tibshirani(1993) and Davison-Hinkley(1997)]. For this reason, we favor Monte Carlo (MC) test methods. MC tests were introduced by Dwass(1957) and Barnard(1963). An important feature of such procedures is that exactness obtains for a given of number of MC replications, without the need to assume that the latter is large or goes to infinity. Further discussions and extensions are also available in [Birnbaum(1974), Dufour-Farhat-Gardiol-Khalaf(1998), Dufour-Khalaf(2001) and Dufour(2006)].

In this paper, we *first* show how the size of all the two-sample homogeneity tests described above can be perfectly controlled for both *continuous* and *discrete* distributions on considering their permutational distribution and using the technique of MC tests prop-

erly adjusted to deal with discrete distributions. As a result, in order to implement these tests, it is not anymore necessary to establish the distributions of the test statistics, either in finite samples or asymptotically.

Second, as a consequence of the great flexibility allowed by the MC test technique in selecting test criteria, we suggest alternative procedures that can provide power gains. These include: (i) a statistic based on the L_∞ distances between kernel-type pdf estimators; (ii) extensions of the permutational test based on the difference of two-sample means to higher order moments, such as sample variances, asymmetry (as third moments) and kurtosis sample coefficients.

Thirdly, on observing that no single test uniformly dominates the others with respect to power, we show that different tests can be combined easily to obtain procedures with better overall power and robustness properties. The procedures proposed involve three steps: (1) in order to make the different statistics comparable, the latter are standardized using first and second moments estimated by simulation; (2) the combined test statistic is defined as the maximum of the standardized test statistics; (3) the MC test technique is used to control the size of a test based on the combined statistic. Depending of the statistics considered different combined tests can be built in this way.

Fourth, we show that the size of these combined test can also be exactly controlled in finite samples through the use of the MC test technique, which will automatically take account of the dependence between the test statistics as well as the discrete nature of their distribution, with a fixed (possibly very small) number of MC replications. It is of interest to note here that such control would be much more difficult, using standard distributional methods, which typically only yield finite-sample (conservative) bounds or large-sample approximations. Typically, combined test procedures are based on the assumption of independence between the test statistics [see the review of Folks(1984)], which does not hold here, or the use of approximations based on bounds [see Miller(1981), Dufour-Torres(1998,2000)] or asymptotic arguments [see Pesarin(2001)]. In contrast, the method we propose for controlling test size does not appeal in any way to the assumption that the number of observations or the number of MC replications go to infinity (as done, for example, in justifying bootstrap techniques).

Fifth, we present the results of a MC experiment which shows clearly that usual large-sample critical values do not control size, while the MC versions of the tests achieve this aim perfectly. Further, we see that the new procedures introduced, either individually or combined with other procedures, can lead to substantial power gains.

In the simulation study, all tests [both the original tests as well as their MC counterparts] were performed at the 5% level using 10000 trials. This entails that the 95%

Table 1: Continuous distributions with their means and variances

Distribution	$N(0,1)$	$Exp(0,1.5)$	$\Gamma(2,1)$	$B(2,3)$	$Log(-1,1)$	$\Lambda(4,\ 1.5)$	$U(0,1)$
Mean	0	1.50	2	.40	-1	168.17	.50
Variance	1	2.25	2	.04	0.55133^{-2}	240055	1/12

confidence interval for the nominal level is [4.57%, 5.43%]. Furthermore, they were all conducted with equal sample sizes $m = n = 22$. As mentioned earlier, each MC test was carried out by picking at random $N = 99$ permutations of the original grouped sample and this was done by using the IMSL (1987) Program Library random number generator. In his simulation study, Allen(1997) used 2500 trials and each permutation or bootstrap test was carried out with 499 samples.

For the first part of the study where F and G are both continuous, the following distributions were considered: normal $N(0,1)$, exponential $Exp(0,\ 1.5)$, gamma $\Gamma(2,\ 1)$, beta $B(2,\ 3)$, logistic $Log(-1,\ 1)$, lognormal $\Lambda(4,\ 1.5)$ and uniform $U(0,\ 1)$. In this choice, care was taken to have at the same time simple parameters as well as appreciably different means and variances. Table 1 gives the list of those means and variances. Four types of situations were considered: (i) the distributions were standardized, and thus had common zero mean and unit variance; (ii) the distributions were only centered, and thus had the zero mean but different variances; (iii) the distributions were only scaled, and thus had different means and common unit variance; (iv) the distributions remained as is and thus had different means and different variances. Whatever the situation, a null hypothesis is obtained each time F and G share the same distribution from the list and an alternative hypothesis is obtained each time F and G possess different distributions from that list.

For the second part of the study where F and G are discrete, the five most commonly used distributions were retained: discrete uniform $[DU(n)]$ on the integers $\{1,\ 2\ ,\ \ldots\ ,\ n\}$, binomial $[Bin(n,\ p)]$, geometric $[Geo(p)]$, negative binomial $[Nbin(N,\ p)]$ and Poisson $[P(\lambda)]$. Since it is a prohibitive task to find parameters that will simultaneously give rise to either common mean and common variance, the following three situations were considered: (i) the distributions were $DU(19)$, $Bin(20,\ 0.5)$, $Geo(0.1)$, $Nbin(8,\ 0.2)$, $P(10)$ and, thus had common mean 10 and variance 30, 5, 90, 2.5 and 10 respectively; (ii) the distributions were $DU(10)$, $Bin(33,\ 0.5)$, $Geo\big((\sqrt{34}-1)/16.5\big)$, $Nbin\big(3,\ (\sqrt{108}-3)/16.5\big)$, $P(8.25)$ and, thus had mean 5.5, 16.5, 3.42, 2.23 and 8.25 respectively but common variance 8.25; (iii) the distributions were $DU(10)$, $Bin\big(10,\ 0.1\big)$, $Geo(0.3)$, $Nbin\big(10,\ 0.2\big)$, $P(5)$ and, thus had mean 5.5, 1, 3.33, 50 and 5 respectively and variance 8.25, 0.9, 7.78, 200 and 5 respectively.

As a check on the accuracy of our study, Tables 1 and 2 of Allen(1997) were reproduced adding, however, the CM, the $\hat{L}_\infty$ and the combined MC tests and by excluding the bootstrap tests.

Most statistics described in the preceding sections have not been well tabulated, so a study of the reliability of tabulated critical values can only be limited. We present some results on this issue for the KS and CM tests. For continuous distributions, we see that the standard KS and CM tests satisfy the level constraint, although the rejection frequencies of the KS test are in some cases notably lower than the level. This can be explained by the fact the 0.05 level cannot be achieved by a non-randomized procedure (due to the discrete character of the distribution), so that the critical values used correspond to smaller sizes. In the case of discrete distributions, it is of interest to note that the KS test can be quite conservative (as predicted by earlier theoretical results), while the CM test can substantially overreject: the CM test is not generally conservative for discrete distributions. In all cases, irrespective of whether the distributions are continuous or discrete, the permutational MC tests have rejection frequencies essentially identical to their nominal levels (as expected).

In this paper, we first showed that finite-sample distribution-free two-sample homogeneity tests, for both continuous and discrete distributions, can be easily obtained on combining two techniques: (1) by considering permutational versions of most proposed tests for that problem; (2) by implementing the permutation procedures as Monte Carlo tests with an appropriate tie-breaking technique to take account of the discreteness of the test null distributions. Second, due to the flexibility of the Monte Carlo test technique, we could easily introduce and implement several alternative procedures, including permutation tests comparing higher-order moments and procedures based on combining several test statistics. Thirdly, in a simulation study, it was shown that the procedures proposed work as expected form the viewpoint of size control, while the new test statistics suggested yield power gains.

Bibliographie

[1] ALLEN, D. L. (1997): Hypothesis Testing Using an L1-Distance Bootstrap, The American Statistician, 51, 145150.
[2] ANDERSON, T. W. (1962): On the Distribution of the Two-Sample Cramr-von Mises Criterion, Annals of Mathematical Statistics, 33, 11481159.
[3] BARNARD, G. A. (1963): Comment on The Spectral Analysis of Point Processes by M. S. Bartlett, Journal of the Royal Statistical Society, Series B, 25, 294.
[4] BIRNBAUM, Z. W. (1974): Computers and Unconventional Test-Statistics, in Reliability and Biometry, ed. by F. Proschan, and R. J. Serfling, pp. 441458. SIAM, Philadelphia, PA.

[5] DAVISON, A., AND D. HINKLEY (1997): Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (UK).
[6] DUFOUR, J.-M. (2006): Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics, Journal of Econometrics, 133, 443477.
[7] DUFOUR, J.-M., A. FARHAT, L. GARDIOL, AND L. KHALAF (1998): Simulation-Based Finite Sample Normality Tests in Linear Regressions, The Econometrics Journal, 1, 154173.
[8] DUFOUR, J.-M., AND M. HALLIN (1993): Improved Eaton Bounds for Linear Combinations of Bounded Random Variables, with Statistical Applications, Journal of the American Statistical Association, 88, 10261033.
[9] DUFOUR, J.-M., AND L. KHALAF (2001): Monte Carlo Test Methods in Econometrics, in Companion to Theoretical Econometrics, ed. by B. Baltagi, Blackwell Companions to Contemporary Economics, chap. 23, pp. 494519. Basil Blackwell, Oxford, U.K.
[10] DUFOUR, J.-M., AND O. TORRS (1998): Union-Intersection and Sample-Split Methods in Econometrics with Applications to SURE and MA Models, in Handbook of Applied Economic Statistics, ed. by D. E. A. Giles, and A. Ullah, pp. 465505. Marcel Dekker, New York.
[11] DUFOUR, J.-M., AND O. TORRS (2000): Markovian Processes, Two-Sided Autoregressions and Exact Inference for Stationary and Nonstationary Autoregressive Processes, Journal of Econometrics, 99, 255289.
[12] DWASS, M. (1957): Modified Randomization Tests for Nonparametric Hypotheses, Annals of Mathematical Statistics, 28, 181187.
[13] EFRON, B., AND R. J. TIBSHIRANI (1993): An Introduction to the Bootstrap, vol. 57 of Monographs on Statistics and Applied Probability. Chapman Hall, New York.
[14] FISZ, M. (1960): On a Result by M. Rosenblatt Concerning the Mises-Smirnov Test, Annals of Mathematical Statistics, 31, 427429.
[15] FOLKS, J. L. (1984): Combination of Independent Tests, in Handbook of Statistics 4: Nonparametric Methods, ed. by P. R. Krishnaiah, and P. K. Sen, pp. 113121. North-Holland, Amsterdam.
[16] LEHMANN, E. L. (1951): Consistency and Unbiasedness of Certain Nonparametric Tests, Annals of Mathematical Statistics, 22, 165179.
[17] MILLER, JR., R. G. (1981): Simultaneous Statistical Inference. Springer-Verlag, New York, second edn.
[18] PESARIN, F. (2001): Multivariate Permutation Tests with Applications in Biostatistics. John Wiley Sons, New York.
[19] ROSENBLATT, M. (1952): Limit Theorems Associated with Variants of the von Mises Statistic, Annals of Mathematical Statistics, 23, 617623.
[21] SMIRNOV, N. V. (1948): Table for Estimating the Goodness of Fit of Empirical Distributions, Annals of Mathematical Statistics, 19, 279281.

Sur la convergence asymptotique des M-estimateurs pondérés

Mohammed EL ASRI

Université d'Avignon
Laboratoire d'Analyse Non Linéaire et Géométrie
33 rue Louis Pasteur 84000 Avignon

Résumé

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ issue de la même loi paramétrique P_θ, avec θ dans un ouvert Θ de $\mathbb{R}^d$. Soit $\rho(x,\theta)$ une fonction mesurable en x, et telle que $\theta_0 := \underset{\theta\in\Theta}{\operatorname{argmin}} E(\rho(X,\theta)) = \underset{\theta\in\Theta}{\operatorname{argmin}} \int \rho(x,\theta) dP_\theta$ existe et soit unique. Dans cet article, nous étudions la classe des M-estimateurs pondérés $\hat{\theta}_n(X_1, X_2, \ldots, X_n) := \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \omega_i \rho(X_i,\theta)$ pour estimer θ_0. Nous établissons la consistance et la normalité asymptotique de $\hat{\theta}_n$. Cette classe engendre, entre autres, l'estimateur de la médiane spatiale pondérée étudiée par Nevalainen et al. (2007). Nous détaillons cet exemple ainsi que celui de l'estimateur de Huber pondéré.

Abstract

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables i.i.d in $\mathbb{R}^d$ with the same parametric distribution P_θ with parameter $\theta \in \Theta$ an open subset of $\mathbb{R}^d$. Let $\rho(x,\theta)$ a x-measurable function such that there exist an unique θ_0 defined by $\theta_0 := \underset{\theta\in\Theta}{\operatorname{argmin}} E(\rho(X,\theta)) = \underset{\theta\in\Theta}{\operatorname{argmin}} \int \rho(x,\theta) dP_\theta$. In this paper, we use the class of weighted M-estimators $\hat{\theta}_n := \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \omega_i \rho(X_i,\theta)$ to estimate θ_0. We establish consistency and asymptotic normality of $\hat{\theta}_n$. A special case of this estimator class, the weighted spatial median, is presented in Nevalainen et al. (2007). We detail this example as well as the weighted Huber estimator.

1 Introduction

La classe des M-estimateurs est introduite par Huber en 1964. Cette classe engendre des estimateurs tels que la médiane, la moyenne empirique et l'estimateur du maximum de vraisemblance. Sur le sujet, on trouve notamment dans les articles d'Huber (1981), Hampel et al. (1986), Van Der Vaart (2000), Van de Geer (2009), Massé et Plante (2003) et Arcones (1998).
Dans le cas de l'estimation d'un paramètre de localisation, la médiane spatiale peut

présenter de très bonnes qualités de robustesse (avec un point de rupture maximal) et de bonnes propriétés de convergence asymptotique. Par exemple, Brown (1983) et Chakraborty et Chaudhuri (1998) ont étudié ces propriétés dans le cas i.i.d.
Nevalainen et al. (2006) et (2007) ont étudié, dans le cas des variables clusterisées, deux types de médianes spatiales : pondérée et non-pondérée. Ils montrent la consistance, la normalité asymptotique de leurs estimateurs et le fait que la pondération permette d'améliorer l'efficacité relative de la médiane spatiale. Le résultat reste valable même dans le cas i.i.d où la corrélation intracluster est nulle.

Il paraît pertinent, de généraliser ces résultats pour la classe des M-estimateurs pondérés. Dans un premier temps, nous explicitons les hypothèses qui permettent d'étudier le comportement asymptotique des M-estimateurs pondérés (consistance et normalité). Enfin, nous détaillons deux exemples d'application : la médiane spatiale et l'estimateur de Huber pondérés.

2 Résultats théoriques

Définitions et notations

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ de loi paramétrique P_θ, $\theta \in \Theta$ où Θ est un ouvert de $\mathbb{R}^d$. Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires i.i.d dans $\mathbb{R}^d$ issue de la même loi P_θ de X. On suppose l'existence et l'unicité de $\theta_0 \in \Theta$ défini par $\underset{\theta\in\Theta}{\operatorname{argmin}} E(\rho(X,\theta)) = \underset{\theta\in\Theta}{\operatorname{argmin}} \int \rho(x,\theta) dP_\theta$, où pour tout $\theta \in \Theta$, $\rho(x,.)$ est une fonction de $\mathbb{R}^d \to \mathbb{R}$ supposée mesurable en x et différentiable dans un voisinage de θ_0. Pour $\psi = \frac{\partial \rho}{\partial \theta}$, alors θ_0 vérifie $E(\psi(X,\theta_0)) = 0$.
Un M-estimateur pondéré associé à la fonction ρ est alors défini par :

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \omega_i \rho(X_i, \theta)$$

où les ω_i sont des poids positifs choisis par le statisticien.
La différentiabilité de $\rho(\cdot,\theta)$ en θ entraîne alors que $\hat{\theta}_n$ est aussi la valeur de θ vérifiant :

$$\sum_{i=1}^{n} \omega_i \psi(X_i, \theta) = 0$$

Hypothèses 1
H_1) Pour tout $\epsilon > 0$, θ_0 vérifie:

$$\inf_{\|\theta-\theta_0\|>\epsilon} E\rho(X,\theta) > E\rho(X,\theta_0)$$

H_2)*a)* $\frac{1}{n}\sum_{i=1}^{n} w_i \to 1$

b) $\lim_n \frac{1}{n}\sum_{i=1}^{n} \omega_i^2 = c_\omega$, *avec* $c_\omega < \infty$.

c) $\lim_n \sum_{i=1}^{n} \frac{\omega_i^2}{i^2} < \infty$.

H_3) *Il existe une fonction* k *et un réel* η *tels que:*

a) $E(k(X)^{2+\eta}) < \infty$, *et* $\forall \theta_1, \theta_2 \in \Theta$

$$|\rho(x,\theta_1) - \rho(x,\theta_2)| \leq k(x)\,\|\theta_1 - \theta_2\| \qquad x - p.p$$

b) $E(\|\psi(X,\theta_0)\|^{2+\eta}) < \infty$, *où* $\|.\|$ *est la norme euclidienne de* $\mathbb{R}^d$

c) $\lim_n \frac{1}{n}\sum_{i=1}^{n} \omega_i^{2+\eta} < \infty$

H_4) *Pour tout* θ *dans un voisinage de* θ_0, *on a:*

$$E\left(\rho(X,\theta)\right) - E\left(\rho(X,\theta_0)\right) = 1/2(\theta-\theta_0)^T V_{\theta_0}(\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$$

avec $V_{\theta_0} = E(\frac{\partial\psi(X,\theta)}{\partial\theta}/_{\theta=\theta_0})$ *inversible.*

Théorème 1

i) Sous les hypothèses H_1, H_2 *et* H_3a): $\hat{\theta}_n \xrightarrow{p} \theta_0$.

ii) Si on vérifie, en plus, les hypothèses H_3b) *et* H_4, *alors*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathrm{N}\left(0, c_\omega V_{\theta_0}^{-1} E\psi(X,\theta_0)\psi^T(X,\theta_0) V_{\theta_0}^{-1}\right)$$

Dans la littérature, on trouve des hypothèses de même type que celles énoncées dans le théorème. Par exemple, dans Van Der Vaart (2000), on retrouve l'hypothèse H_1 qui traduit l'unicité de θ_0, ainsi que l'hypothèse H_3 qui définit la condition de Lipschitz de la fonction ρ. Si les poids $\omega_i, i = 1, 2, \ldots$ sont uniformement bornés, alors l'hypothèse H_2 sont automatiquement vérifées. Dans Arcones (1998), l'hypothèse H_4 est montrée pour des fonctions $\rho(x,\theta) = f\left(\|x-\theta\|\right)$, où f est une fonction réelle, convexe, avec une dérivée première continue et une dérivée seconde positive, bornée et uniformément continue; ceci est le cas, par exemple, si $f(a) = a, a \in \mathbb{R}$ (cas de la médiane spatiale).

Éléments de preuve

i) On cherche à établir, dans un premier temps, la convergence en probabilité de $\hat{\theta}_n$ vers θ_0. On note $M(\theta) = E(\rho(X,\theta))$ et $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \rho(X_i,\theta)$. Cette étude est développée

en 3 étapes:
Étape 1 : On cherche à établir la limite suivante pour tout θ: $M_n(\theta) \xrightarrow{p.s} M(\theta)$, en vérifiant le critère de Kolmogorov [Serfling (1980) p.27].
Étape 2 : On exploite le résultat de l'étape 1 et on utilise les conditions de Lipschitz des fonctions M_n et M, pour établir la limite: $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s} 0$.
Étape 3 : En utilisant les résultats des deux étapes précédentes et la définition de $\hat{\theta}_n$, on montre que $\hat{\theta}_n \xrightarrow{p} \theta_0$. On conclut par cette étape la démonstration de i).

ii) On démontre la normalité asymptotique de $\hat{\theta}_n$ en deux étapes principales :
Étape 1 : On introduit la statistique $T_n = \frac{1}{n}\sum_{i=1}^{n} \omega_i \psi(X_i, \theta_0)$, dont on établit la normalité asymptotique, en vérifiant les hypothèses du théorème p.23-31 Serfling (1980).
Étape 2 : On établit la relation liant T_n avec notre estimateur $\hat{\theta}_n$, en se basant principalement sur l'hyphothèse H_4 : $\left\| V_{\theta_0}^{1/2}(\hat{\theta}_n - \theta_0) + V_{\theta_0}^{-1/2} T_n \right\| = o_p(1)$.
On vérifie par ces deux étapes, les hypothèses du théorème p.10 Van Der Vaart (2000) et on montre la normalité asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

3 Exemples

Exemple 1: La médiane spatiale pondérée

La médiane spatiale pondérée $\hat{\theta}_n$ est la valeur de θ qui minimise la fonction objective $\frac{1}{n}\sum_{i=1}^{n} \omega_i \|X_i - \theta\|$. On suppose que le support de la loi P_θ n'est pas concentrée sur une ligne. Ainsi, d'après Milasevic et Ducharme (1987), θ_0 est unique. La fonction $\rho(x, \theta) = \|x - \theta\|$ est 1-Lipschitz donc elle vérifie l'hypothèse H_2 avec $k(x) \equiv 1$; sa dérivée en θ est donnée par : $\psi(x, \theta) = \frac{(x-\theta)}{\|x-\theta\|} \quad \forall x \neq \theta$. Comme $\|\psi\| = 1$, la condition H_3 est également vérifiée. La fonction ρ s'écrit également: $\rho(x, \theta) = f(\|x - \theta\|)$, avec $f(a) = a$ et a dans $\mathbb{R}$. D'après l'article d'Arcones (1998), la fonction f remplit les conditions pour que l'hypothèse H_4 soit vérifiée. Si en plus l'hypothèse H_2 est vérifiée, alors $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converge vers une loi normale de moyenne 0 et de variance: $V_{\theta_0}^{-1} c_\omega E \frac{(X-\theta_0)(X-\theta_0)^T}{\|X-\theta_0\|^2} V_{\theta_0}^{-1}$ avec $V_{\theta_0} = 2I_{d\times d}$.

Exemple 2: La fonction d'Huber

La fonction ρ de Huber est définie par :

$$\rho(x, \theta) = \begin{cases} \frac{1}{2}\|x - \theta\|^2 & \text{si } \|x - \theta\| \leq k \\ k\|x - \theta\| - \frac{1}{2}k^2 & \text{si } \|x - \theta\| > k, \end{cases}$$

Pour vérifier l'hypothèse H_1, on suppose que la fonction $\theta \mapsto E(\rho(X, \theta))$ admet un minimum unique θ_0. La fonction d'Huber est k-Lipschizienne. Elle vérifie donc l'hypothèse H_3. L'hypothèse H_4 est également vérifiée. En effet, la fonction d'Huber s'écrit comme

une fonction de $\|x-\theta\|$ qui vérifie les conditions suffisantes de l'article d'Arcones (1998). On peut donc appliquer notre résultat pour établir la normalité asymptotique de l'estimateur avec une moyenne nulle et une variance $V_{\theta_0}^{-1} c_\omega D_{\theta_0} V_{\theta_0}^{-1}$ où

$V_{\theta_0} = E(2I_{d\times d} 1_{\|X-\theta_0\|\leq k}$
$+k\left[-\|X-\theta_0\|^{-3}(X-\theta_0)(X-\theta_0)^T + \|X-\theta_0\|^{-1} I_{d\times d}\right] 1_{\|X-\theta_0\|>k})$
et $D_{\theta_0} = E(\left[(X-\theta_0)(X-\theta_0)^T\right] 1_{\|X-\theta_0\|\leq k}$
$+k^2\left[-\|X-\theta_0\|^{-2}(X-\theta_0)(X-\theta_0)^T\right] 1_{\|X-\theta_0\|>k})$

4 Discussion

On a étudié les propriétés asymptotiques des M-estimateurs pondérés en établissant leur convergence en probabilité et leur normalité asymptotique. On a ainsi une variance asymptotique de l'estimateur qui dépend des poids mais qui n'améliore pas son efficacité (l'efficacité optimale est obtenue avec des poids $\omega_i \equiv 1$, cas non-pondéré). Notre travail actuel consiste à étudier la robustesse de ces estimateurs, en particulier leur point de rupture.
Dans un cas particulier, Nevalainen et al. (2007) étudient la médiane spatiale pondérée pour des variables clusterisées. En plus de l'efficacité, ils montrent qu'elle est moins sensible à la corrélation intracluster que la médiane spatiale non-pondérée. Ces éléments nous permettent d'envisager un élargissement du spectre de la recherche autour des M-estimateurs pondérés en analysant les propriétés sur d'autres types de variables, notamment, les variables clusterisées.

Bibliographie

[1] Arcones, M. A. (1998) Asymptotic theory for M-estimators over a convex kernel, Econometric Theory 14, n°. 4: 387-422.
[2] Brown, B. M. (1983) Statistical uses of the spatial median, Journal of the Royal Statistical Society, Series B (Methodological) 45, n°. 1: 25-30.
[3] Chakraborty, B., et P. Chaudhuri (1998) On an adaptive transformation-retransformation estimate of multivariate location, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60, n°. 1: 145-157.
[4] Hampel F. R., E. M. Ronchetti, P. J. Rousseeuw, et W. A. Stahel (1986) Robust statistics : the approach based on influence functions, New York: Wiley.
[5] Huber, P. (1981) Robust statistics, New York: Wiley.
[6] Massé, J., et J. Plante (2003) A Monte Carlo study of the accuracy and robustness of ten bivariate location estimators, Computational Statistics & Data Analysis 42 : 1-26.
[7] Milasevic, P., et G. H. Ducharme (1987) Uniqueness of the Spatial Median, The Annals of Statistics 15, n°. 3 (Septembre): 1332-1333.

[8] Nevalainen, J., D. Larocque, et H. Oja (2006) A weighted spatial median for clustered data. Statistical Methods and Applications 15, n°. 3 (11): 355-379.
[9] Nevalainen, J., D. Larocque, et H. Oja (2007) On the multivariate spatial median for clustered data, Canadian Journal of Statistics 35, n°. 2 (6): 215-231.
[10] Rockafellar, R. Tyrrell (1997) Convex analysis, Princeton University Press.
[11] Serfling, R. (1980) Approximation theorems of mathematical statistics, New York: Wiley.
[12] Van De Geer, S. (2009) Empirical processes in M-estimation, Cambridge: Cambridge University Press.
[13] Van Der Vaart, A. (2000) Asymptotic statistics, Cambridge: Cambridge University Press.

Vraisemblance empirique pour les séries temporelles périodiques

Hugo Harari-Kermadec & Jacek Leśkow

Hugo Harari-Kermadec, ENS Cachan et Laboratoire SAMM, Université Paris 1, 90, rue de Tolbiac, 75634 PARIS CEDEX 13
Jacek Leśkow, Wyższa Szkoła Biznesu WSB-NLU Nowy Sącz

Mots clés : Méthodes semi-paramétriques; PARMA; non stationnarité.

Résumé :

Dans cette présentation, nous proposons d'utiliser la vraisemblance empirique pour des données dépendantes suivant un modèle périodique. L'originalité de cette étude tient à la non stationnarité des séries étudiées. Pour simplifier la présentation, nous étudierons le cas d'un processus auto-régressif périodique d'ordre 1, un P-AR(1). La construction de blocs de données respectant la structure de dépendance est centrale ici : les données étant P périodique, nous proposons de choisir des blocs de longueur proportionnelle à P.

Abstract:

In this paper, we propose a method based on empirical likelihood for cyclic time series. The innovation of the paper comes from the nonstationnarity of cyclic time series. In order to simplify the presentation, we study the case of a periodic autoregressive time series of order 1, namely a P-AR(1). A key issue here is the construction of data blocs following the dependence structure : since data is P periodic, we introduce blocs with length proportional to P.

1 Introduction

Les séries temporelles présentant une espérance ou une structure de covariance périodiques sont d'usage courant dans différents domaine d'application, en traitement du signal, en climatologie ou en vibro-méchanique. Les séries économiques ou financières peuvent également présenter des comportements saisonniers. Pour une revue de la littérature des applications en traitement du signal, voir Gardner et al (2006).

La périodicité est une forme contrôlée de non stationnarité qui conduit néanmoins à d'importantes complication dans les preuves de convergence des méthodes usuelles. La principale difficulté provient de la difficulté à écrire la matrice de variance-covariance des estimateurs. Des méthodes de type Bootstrap sont donc employées pour contourner cette difficulté et estimer de façon non paramétrique cette variance-covariance.

Nous proposons ici d'utiliser la vraisemblance empirique qui possède, en plus de similarités fortes avec le Bootstrap, une très appréciable propriété d'auto-normalisation. La matrice de variance-covariance des estimateurs est en quelque sorte prise en compte "en

interne" de la procédure d'estimation et on obtient simplement des tests ou des régions de confiance.

Pour traiter de données dépendantes, les méthodes semi-paramétriques sont appliquées à des blocs de données plutôt qu'aux données elles-même. Politis (2003) propose une revue de la littérature sur le Bootstrap pour données dépendantes. Ces constructions ont été reprises dans le cadre de la vraisemblance empirique, voir Kitamura (1997) ou Harari-Kermadec (2011) par exemple.

2 Le modèle périodique auto-régressif d'ordre 1

Les modèles périodiques correspondent à des modèles classiques dont les paramètres varient de façon périodique. Le plus simple de ces modèles est sans doute le P-AR(1). Soit $(X_t)_{t\in\mathbb{Z}}$ une série telle que

$$X_{jP+\nu} - \phi(\nu)X_{jP+\nu-1} = \sigma_\nu \varepsilon_{jP+\nu} \ , \nu = 1, \dots, P, \tag{1}$$

où P est la période et les résidus ε_t sont supposés indépendants et identiquement distribués (i.i.d.) de variance 1. Bien qu'il s'agisse d'un P-AR(1), il y a $2P$ paramètres : $(\phi(1), \dots, \phi(P), \sigma_1^2, \dots, \sigma_P^2)$, avec $\theta_0 = (\phi(1), \dots, \phi(P))$ la vraie valeur du paramètre P-dimensionnel d'intérêt et $(\sigma_1^2, \dots, \sigma_P^2)$ les paramètres de nuisance.

L'équation (1) permet de définir les paramètres d'intérêt à l'aide d'un système d'équations de moments comme c'est usuellement le cas dans la méthode classique de vraisemblance empirique. On peut l'écrire sous forme vectoriel pour faire disparaître ν :

$$\mathbb{E}\left[\begin{pmatrix} X_{jP+P} \\ \cdots \\ X_{jP+1} \end{pmatrix} - \theta_0 \begin{pmatrix} X_{jP+P-1} \\ \cdots \\ X_{jP} \end{pmatrix}\right] = \mathbb{E}\left[\begin{pmatrix} \sigma_P \varepsilon_{jP+P} \\ \cdots \\ \sigma_1 \varepsilon_{jP+1} \end{pmatrix}\right] = 0$$

ou

$$\mathbb{E}\left[\mathbf{X}_j - \theta_0 L^{\otimes P} \mathbf{X}_j\right] = 0$$

où $\mathbf{X}_j = (X_{jP+1}, \dots, X_{jP+P})'$ et $L^{\otimes P}$ est l'opérateur retard pour un vecteur de P séries temporelles.

3 Vraisemblance empirique

Dans le cas particulièrement simple d'un P-AR(1), les vecteurs $\mathbf{X}_j - \theta_0 L^{\otimes P}\mathbf{X}_j$ sont i.i.d. . On peut appliquer directement la vraisemblance empirique classique à ces vecteurs de taille P pour estimer les paramètre d'intérêt. Cette simplicité provient de l'égalité entre le nombre de paramètres d'intérêt et le nombre d'équation de moments, et ne se généralise pas directement aux modèles plus complets.

Soit n le nombre d'observations et m la partie entière de n/P. Le rapport de vraisemblance empirique r_n s'écrit :

$$r_n(\theta) = -\sup_{(q_1,\cdots,q_m)\in[0,1]^m} \left\{ \log\left(\prod_{j=1}^{m} mq_j\right) \middle| \sum_{j=1}^{m} q_j \left(\mathbf{X}_j - \theta L^{\otimes P}\mathbf{X}_j\right) = 0,\ \sum_{j=1}^{m} q_j = 1 \right\}.$$

En utilisant un Lagrangien, on peut réécrire cette optimisation :

$$r_n(\theta) = \sup_{\lambda\in\mathbb{R}^P} \left\{ \sum_{j=1}^{m} \log\left[1 + \lambda'\left(\mathbf{X}_j - \theta L^{\otimes P}\mathbf{X}_j\right)\right] \right\}.$$

Cette écriture permet de souligner l'indépendance bien connue de la vraisemblance par rapport à l'échelle : on normalise le problème en posant

$$\gamma = \lambda' I_P \begin{pmatrix} \sigma_P \\ \cdots \\ \sigma_1 \end{pmatrix}.$$

Le théorème suivant est un corolaire direct des résultats classiques de la vraisemblance empirique (Owen, 1990).

Théorème 1 *Soit $(X_1,\ldots,X_n)$ un P-AR(1) tel que pour $\theta_0 = (\phi(1),\ldots,\phi(P))$ et*

$$\forall \nu \in [|1;P|],\ X_{jP+\nu} - \phi(\nu) X_{jP+\nu-1} = \sigma_\nu \varepsilon_{jP+\nu}\ ,\ \textit{avec } \sigma_\nu > 0$$

où les ε_t sont i.i.d., centrés et de variance 1. L'estimateur du maximum de vraisemblance empirique

$$\hat{\theta} = (\hat{\phi}(1),\ldots,\hat{\phi}(P)) = \arg\inf_{\theta\in\mathbb{R}^P} r_n(\theta)$$

est un estimateur asymptotiquement gaussien de θ. De plus,

$$2r_n(\theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_P$$

et donc

$$C_n = \left\{ \theta \in \mathbb{R}^P,\ 2r_n(\theta) \leq F^{-1}_{\chi^2_P}(1-\alpha) \right\}$$

est une région de confiance asymptotique de niveau $1-\alpha$ pour θ_0.

4 Exemple illustratif

Nous proposons ici une illustration très simple de notre méthode pour un 2-AR(1). Soit $(X_1,\ldots,X_n)$ tel que $X_1 = \varepsilon_1$ et

$$\forall t \geq 1,\ X_{t+1} - \sin\left(\frac{2\pi t + 1}{2}\right) X_t = \varepsilon_t,$$

où les ε_t sont i.i.d. uniformes, centrés et de variance $1/12$.

On obtient les régions de confiance suivantes :

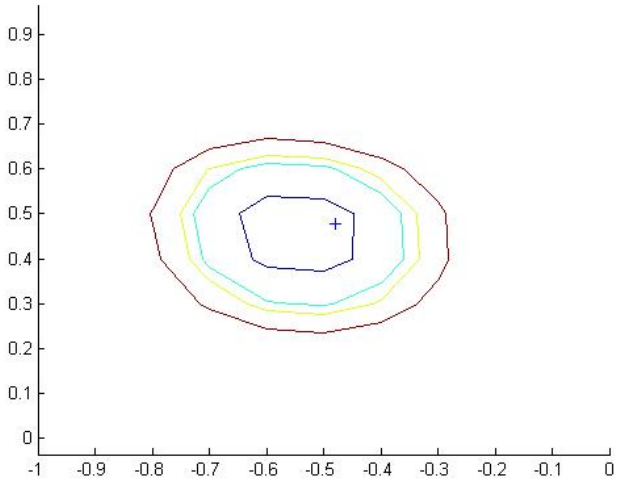


Figure 1: Zones de confiance de niveau 50%, 90%, 95% et 99% pour un 2-AR(1) avec innovations uniformes. La croix indique la vraie valeur du paramètre θ_0.

5 Perspective

L'objectif de ce travail préliminaire est d'étendre cette méthode aux modèles ARMA périodiques (P-ARMA) et au-delà aux données faiblement dépendantes non stationnaires presque périodiques. Pour ce faire, le découpage en bloc introduit dans le cadre de la vraisemblance empirique par Kitamura (1997) associé à des conditions d'α-mélangeance devrait permettre d'obtenir des résultats de convergence satisfaisant.

Nous envisageons de généraliser cette méthode à la famille est des modèles P-ARMA et plus largement aux données faiblement dépendantes P-périodiques en distribution :

$$\forall (t_1, \ldots, t_a, b) \in \mathbb{Z}^{a+1},\ (X_{t_1}, \ldots, X_{t_a}) \sim (X_{t_1+bP}, \ldots, X_{t_a+bP})$$

On se donne une série faiblement dépendante P-périodiques en distribution $(X_t)_{t\in\mathbb{Z}}$ et un vecteur de paramètres d'intérêt $\theta_0 \in \mathbb{R}^k$ défini une équation de moment

$$M(\theta_0, X_{t-r}, \ldots, X_{t+P}) = 0 \tag{2}$$

où P est la période et r le nombre de retard nécessaires à la définition des paramètres (typiquement l'ordre de la partie AR).

Pour adapter la vraisemblance empirique à la dépendance faible, Kitamura (1997) propose de reprendre un méthode issue de la littérature du Bootstrap. Pour tenir compte de la périodicité, nous proposons de choisir des longueur de blocs proportionnelles à la période P : Soit BP la longueur des blocs et SP la séparation entre deux débuts de blocs

consécutifs. Les blocs B_j sont alors donnés par :

$$B_j = (X_{jSP-r}, \ldots, X_{(jS+B)P}) = (X_{jSP-r}, \ldots, X_{jSP}, \mathbf{X}'_{jS}, \ldots, \mathbf{X}'_{(jS+B)}). \tag{3}$$

Pour tout bloc de cette forme et pour toute valeur des paramètres θ, on pose :

$$Y_j^B(\theta) = \frac{1}{B}\sum_{k=0}^{B-1} M(\theta, X_{(jS+k)P-r}, \ldots, X_{(jSP+k)P+P}),$$

la moyenne des réalisations de $M(\theta, \dot{})$ sur les périodes incluses dans le bloc.

Nous conjecturons que sous des hypothèses de régularité sur M et de mélangeance sur (X_t), on obtient un théorème équivalent au Théorème (1) en construisant le rapport de vraisemblance à partir des Y_j^B, B et S de l'ordre de $o(\sqrt{n})$.

Bibliographie

[1] Antoine, B., Bonnal, H. et Renault, E. (2007), On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood, *Journal of Econometrics*, 138, 461–487.

[2] Gardner, W. A., Napolitano, A. et Paura, L. (2006), Cyclostationarity: Half a century of research, *Signal Processing*, 86(4), 639–697.

[3] Harari-Kermadec, H. (2011), Regenerative block empirical likelihood for Markov chains, *Journal of Nonparametrics Statistics*, In press.

[4] Kitamura, Y. (1997), Empirical likelihood methods with weakly dependent processes, *Annals of Statistics*, 25(5), 2084–2102.

[5] Owen, A. B. (1990), Empirical likelihood ratio confidence regions, *Annals of Statistics*, 18, 90–120.

[6] Politis, D. N. (2003), The Impact of Bootstrap Methods on Time Series Analysis, *Statistical Science*, 18(2), 219–230.

A Démonstration du Théorème 1

On reprend ici les lignes de la démonstration de la convergence du rapport de vraisemblance empirique (Owen, 1990). On pose $Y_j = \mathbf{X}_j - \theta_0 L^{\otimes P}\mathbf{X}_j$. On rappelle que

$$Y_j = \begin{pmatrix} X_{jP+P} \\ \cdots \\ X_{jP+1} \end{pmatrix} - \theta_0 \begin{pmatrix} X_{jP+P-1} \\ \cdots \\ X_{jP} \end{pmatrix} = \begin{pmatrix} \sigma_P \varepsilon_{jP+P} \\ \cdots \\ \sigma_1 \varepsilon_{jP+1} \end{pmatrix}$$

est i.i.d. .

Le rapport de vraisemblance empirique $r_n(\theta_0) = \sup_{\lambda \in \mathbb{R}^P} \sum_{j=1}^m \log(1 + \lambda' Y_j)$ avec m la partie entière de n/P. La condition du premier ordre au supremum λ_n est alors :

$$1/m \sum_{j=1}^{m} \frac{Y_j}{1 + \lambda_n' Y_j} = 0. \tag{4}$$

Soit $u = \lambda_n/||\lambda_n||$. Un calcul classique (Owen, 1990) conduit à

$$u'\overline{Y} \geq ||\lambda_n|| \left[u' S_m^2 u - \max_j ||Y_j|| u'\overline{Y} \right] \tag{5}$$

où $S_m^2 = 1/m \sum_{j=1}^m Y_j' Y_j$.

On contrôle le terme entre crochets en plusieurs étapes. Premièrement, par la loi des grands nombres, S_m^2 converge en probabilité vers S^2 la matrice diagonale de variance-covariance des Y_i dont la diagonale est composée des σ_ν^2. On a alors

$$\min_{1\leq\nu\leq P} \sigma_\nu^2 + o(1) \leq u' S_m^2 u \leq \max_{1\leq\nu\leq P} \sigma_\nu^2 + o(1).$$

Deuxièmement, en appliquant le Lemme A.1 de Antoine et al. (2007), on obtient $\max_j ||Y_j|| = o(n^{1/2})$. Enfin, le théorème central limite appliqué au Y_j donne $\overline{Y} = \mathcal{O}(n^{-1/2})$. L'inegalité (5) donne alors

$$\mathcal{O}(n^{-1/2}) \geq ||\lambda_n|| \left[u' S_m^2 u - o(n^{1/2}) \mathcal{O}(n^{-1/2}) \right] = ||\lambda_n|| (u' S_m^2 u + o(1)),$$

et $||\lambda_n||$ is then $\mathcal{O}(n^{-1/2})$.

La condition du première ordre (4) et l'égalité classique $1/(1+x) = 1 - x + x^2/(1+x)$, permet d'obtenir

$$0 = 1/m \sum_{j=1}^m Y_j \left(1 - \lambda_n' Y_j + \frac{(\lambda_n' Y_j)^2}{1 + \lambda_n' Y_j} \right) = \overline{Y} - S_m^2 \lambda_n + 1/m \sum_{j=1}^m \frac{Y_j (\lambda_n' Y_j)^2}{1 + \lambda_n' Y_j}.$$

Le dernier terme est d'ordre $o(n^{-1/2})$ d'après le Lemme A.2 de Antoine et al. (2007) et donc $\lambda_n = S_m^{-2} \overline{Y} + o(n^{-1/2})$.

On peut alors développer au second ordre le rapport de vraisemblance :

$$2r_n(\theta_0) = 2 \sum_{j=1}^m \log(1 + \lambda_n' Y_j) = 2m \lambda_n' \overline{Y} - m \lambda_n' S_m^2 \lambda_n + 2 \sum_{j=1}^m \eta_j,$$

où les η_j sont tels qu'il existe $B > 0$ et avec une probabilité tendant vers 1, $|\eta_j| \leq B |\lambda_n' Y_j|^3$. On a alors en appliquant à nouveau le Lemme A.1 de Antoine et al. (2007) et en remarquant que m est du même ordre que n,

$$2 \sum_{j=1}^m |\eta_j| \leq \sum_{j=1}^m \|Y_j\|^3 \leq m \max_j ||Y_j|| \left(\frac{1}{m} \sum_{j=1}^m \|Y_j\|^2 \right) = m \; o(m^{1/2}) \mathcal{O}(1) = o(n^{3/2})$$

Finalement,

$$2r_n(\theta_0) = 2m \lambda_n' \overline{Y} - m \lambda_n' S_m^2 \lambda_n + o(1) = m \overline{Y}' S_m^{-2} \overline{Y} + o(1) \xrightarrow[n\to\infty]{\mathcal{L}} \chi_p^2.$$

Biostatistique

Utilisation d'un Modèle Linéaire Mixte Multivarié pour identifier les éléments managériaux et individuels associés à l'appropriation de recommandations professionnelles dans des services de médecine, *Kret Marion, Domecq Sandrine, Saillour Glenisson Florence and Sibe Matthieu*

Le problème de l appropriation des Recommandations Professionnelles (RP) par les professionnels de santé est crucial pour la qualité des soins [1]. L objectif du projet national Theorem était d identifier les caractéristiques managériales et organisationnelles de services de médecine, et les caractéristiques individuelles des professionnels, associées à l'appropriation des RP. Pour cela, plusieurs outils quantitatifs ont été élaborés. Un questionnaire individuel et un questionnaire de culture organisationnelle ont été complétés par chaque professionnel des 36 services sélectionnés par tirage au sort. Des faits organisationnels ont été recueillis au niveau service. Ce plan de sondage a amené à l utilisation de données hiérarchiques [2]. Un modèle linéaire mixte a été utilisé afin de prendre en compte les sources de variation de chaque niveau hiérarchique. Par ailleurs le degré d appropriation des RP a été mesuré au niveau individuel via des scores calculés à partir de cas cliniques portant sur trois thèmes. Afin d apprécier l'appropriation des RP, il était indispensable d étudier conjointement les trois scores. La prise en compte simultanée de données hiérarchiques et de plusieurs variables à expliquer a nécessité l utilisation d'un modèle linéaire mixte multivarié [3]. Il permet d'évaluer l association entre plusieurs variables à expliquer et d étudier des effets globaux ou spécifiques des variables explicatives. Les corrélations entre les trois scores étaient prises en compte dans le modèle. Au total, 17 facteurs ont été retenus parmi 141 potentiels. Certains avaient un effet global sur les scores aux cas cliniques alors que d autres avaient des effets spécifiques sur l un ou l autre des scores.

Application des méthodes bayésiennes à l'analyse d'un événement récurrent, *Anissa Elfakir, Jérôme Tanguy and Sébastien Marque*

Afin de démontrer le bénéfice santé d'un produit alimentaire, il est nécessaire d'établir un plan de développement clinique s'appuyant sur un ensemble cohérent d'études pré-cliniques et cliniques. La statistique bayésienne permet d'intégrer dans l'analyse d'une nouvelle étude les connaissances a priori telles que les avis d'experts, ou les résultats des études précédentes. Cette possibilité de capitaliser sur de l'information disponible et hétérogène permet notamment d'accélérer le processus de décision dans le développement clinique. Ces méthodes semblent d'autant plus attrayantes qu'elles bénéficient ces dernières années d'un contexte réglementaire favorable et sont à présent implémentées dans les logiciels couramment utilisés et acceptés par l'industrie.

Dans le cadre d'une étude clinique en nutrition, l'effet d'un produit laitier sur les occurrences multiples d'un évènement a été analysé en utilisant deux familles de méthodes fréquentistes complémentaires (donnée de comptage et donnée de survie). Cette problématique a également été analysée à travers des méthodes bayésiennes afin d'évaluer la sensibilité des résultats.

L'analyse a été faite à l'aide des nouvelles procédures bayésiennes implémentées dans le logi-

ciel SAS® version 9.2 (SAS Institute, Inc, Cary, NC). Une attention particulière à été portée au concept d'élicitation de connaissances expertes. L'influence de la distribution a priori sur le paramètre d'estimation de l'effet produit a été testée en utilisant plusieurs lois informatives et non informatives. Pour évaluer la qualité de l'estimation des paramètres, divers diagnostiques de convergences usuels des chaines de Markov ont été étudiés (graphiques et indicateurs statistiques). Les possibilités d'interprétation offertes par les méthodes bayésiennes et leur l'intérêt par rapport aux méthodes fréquentistes ont été discutées.

Méthodes de correction du degré de signification pour une recherche de codage optimal dans un modèle linéaire généralisé, *Jérémie Riou and Benoit Liquet*

En modélisation, la recherche d'un codage optimal pour une variable quantitative entraîne la réalisation de nombreux tests, qui nécessitent de corriger le degré de signification. Nous nous placerons tout d'abord dans le cadre d'une codification binaire pour un modèle linéaire généralisé. Dans ce contexte, il existe une correction exacte du degré de signification du test défini comme le maximum des tests associés aux différents codages. Cette correction sera comparée à la procédure de Bonferroni, d'Efron et aux procédures de rééchantillonage par permutation et bootstrap. Enfin, nous nous intéresserons à une transformation de la variable en plus de deux classes. Le degré de signification issu du maximum des tests sera ensuite calculé par des méthodes de rééchantillonage. Ces méthodes seront comparées par simulations et appliquées sur données réelles. Un package R permettant la réalisation de ces méthodes sera de plus proposé.

Etude de comparaison des différentes méthodes de recherche de dose en oncologie, avec prise en compte de toxcités modérées et gradées, *Monia Ezzalfani, Marie-Cécile Le Deley and Sarah Zohar*

Le but des essais de phase I est d'identifier la dose à recommander (DR) pour les évaluations ultérieures. Ces essais séquentiels utilisent classiquement un critère binaire de toxicité, ignorant le détail des grades et la multiplicité possible des toxicités. Ce paradigme, développé pour les molécules cytotoxique, est remis en cause par l'émergence des thérapies ciblées, moins toxiques. Le but de notre travail est de comparer les méthodes d'escalade de doses basées sur un score de toxicité, proposées par Yuan(2007), Ivanova(2009) et Chen(2010). Ces méthodes considérant la toxicité comme une variable quasi-continue se différencient par l'inférence statistique et le schéma d'allocation séquentielle des doses : modélisation de la relation dose-score avec inférence Bayésienne pour Yuan ; escalade de dose basée sur un algorithme pour Ivanova et Chen. Nous avons évalué les performances des ces méthodes par une étude de simulations sous différents scénarios de relation dose-score de toxicité monotone croissante (150 scénarios tirés selon une loi uniforme). Nous avons approximé la mesure de toxicité par une loi normale tronquée, avec plusieurs hypothèses sur la variance. En moyenne, les résultats de ces méthodes présentent une bonne performance. La méthode de Yuan s'avère sensiblement plus performante en termes de pourcentage de sélection correcte de la DR et d'écart entre la dose sélectionnée et la vraie DR. Les performances des différentes méthodes sont très liées aux hypothèses faites sur la variance.

La recherche méthodologique dans ce domaine nécessite une collaboration avec les cliniciens, définissant une mesure de toxicité adéquate, afin de valider la pertinence des méthodes.

Fitting an augmented Bayesian network to improve a complex quantitative microbial risk assessment model from durability studies,

Sophie Ancelet, Clémence Rigaux, Frédéric Carlin, Christophe Nguyen-Thé and Isabelle Albert

The Monte-Carlo simulation approach is traditionally used in food safety quantitative risk assessment. When some data are available, such an approach does not allow "back-calculation" in a Quantitative Microbial Risk Assessment (QMRA) Bayesian Network (BN) to borrow strength across the dependency relationships between the variables and then, efficiently learn about the microbial dynamics of a food-borne pathogen. In such an inverse problem, building an augmented BN from an observational model then performing a Bayesian inference may be more fruitful. We aim to show that it is possible and relevant to infer an augmented BN in the context of (i) a much more complex QMRA model that describes the dynamics from "farm-to-fork" of a heterogeneous population of Bacillus cereus in a courgette purée (ii) additional raw data coming from durability studies and that are not directly linked to one given variable of the QMRA model.

JEUDI 26 MAI 2011, 16h50

Utilisation d'un Modèle Linéaire Mixte Multivarié pour identifier les éléments managériaux et individuels associés à l'appropriation de recommandations professionnelles dans des services de médecine.

Marion KRET[a], Sandrine DOMECQ[a], Matthieu SIBE[b], Florence SAILLOUR-GLENISSON[a].

[a] CCECQA – Comité de Coordination de l'Evaluation Clinique et de la Qualité en Aquitaine.
Hôpital Xavier Arnozan
33600 PESSAC

[b] ISPED - Institut de Santé Publique, d'Epidémiologie et de Développement.
Université Victor Segalen Bordeaux 2
146, rue Léo Saignat
33076 BORDEAUX cedex

RESUME / ABSTRACT

Le problème de l'appropriation des Recommandations Professionnelles (RP) par les professionnels de santé est crucial pour la qualité des soins [1]. L'objectif du projet national *Theorem* était d'identifier les caractéristiques managériales et organisationnelles de services de médecine, et les caractéristiques individuelles des professionnels, associées à l'appropriation des RP. Pour cela, plusieurs outils quantitatifs ont été élaborés. Un questionnaire individuel et un questionnaire de culture organisationnelle ont été complétés par chaque professionnel des 36 services sélectionnés par tirage au sort. Des faits organisationnels ont été recueillis au niveau service. Ce plan de sondage a amené à l'utilisation de données hiérarchiques [2]. Un modèle linéaire mixte a été utilisé afin de prendre en compte les sources de variation de chaque niveau hiérarchique. Par ailleurs le degré d'appropriation des RP a été mesuré au niveau individuel via des scores calculés à partir de cas cliniques portant sur trois thèmes. Afin d'apprécier l'appropriation des RP, il était indispensable d'étudier conjointement les trois scores. La prise en compte simultanée de données hiérarchiques et de plusieurs variables à expliquer a nécessité l'utilisation d'un modèle linéaire mixte multivarié [3]. Il permet d'évaluer l'association entre plusieurs variables à expliquer et d'étudier des effets globaux ou spécifiques des variables explicatives. Les corrélations entre les trois scores étaient prises en compte dans le modèle. Au total, 17 facteurs ont été retenus parmi 141 potentiels. Certains avaient un effet global sur les scores aux cas cliniques alors que d'autres avaient des effets spécifiques sur l'un ou l'autre des scores.

Mots clés : *Modèle linéaire mixte multivarié, Données hiérarchiques, Recommandations professionnelles, Services de médecine.*

The issue of guidelines adherence of caregivers and its implementation is very important for the quality of care [1]. The objective of the national project *Theorem* was to examine how guidelines implementation is associated with the managerial and organizational characteristics of medical wards, and with the individual characteristics of caregivers. For that purpose, several quantitative tools were developed. An individual questionnaire and an organisational culture questionnaire were completed by every professional. Organisational facts were collected at the ward's level. This sampling design led to the use of hierarchical data [2]. A linear mixed model was used to take into consider the sources of variation of each hierarchical level. Besides the level of guidelines implementation was measured at the individual level through scores calculated from clinical cases on three themes. In order to assess guidelines implementation, it was necessary to study jointly the three scores. The simultaneous consideration of hierarchical data and several dependent variables required the use of a multivariate linear mixed model [3]. It allows to estimate the association between several variables to explain and to study global or specific effects of the independent variables. The correlations between three scores were taken into account in the model. In the end, 17 factors were selected among 141 potential. Some had a global effect on the scores of the three scores, while others had specific effects one or other of the scores.

Keywords: *Multivariate linear mixed models, Hierarchical data, Guidelines implementation, Medicine wards.*

BIBLIOGRAPHIE

[1] Schouten JA, Hulscher ME, Kullberg BJ, Cox A, Gyssens IC, Van der Meer JW, Grol RP (2005) Understanding variation in quality of antibiotic use for community-acquired pneumonia: effect of patient professional and hospital factors. *J Antimicrob Chemother,* 56 : 575-82.

[2] Arrègle JL (2003). Les modèles linéaires hiérarchiques : principe et illustration. *Management*, 6(1): 1-28.

[3] Thiébaut R, Jacqmin-Gadda H, Chene G, Leport C and Commenges D (2002). Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine*, 69: 249-256.

INTRODUCTION

Face à la politique actuelle d'élaboration de Recommandations Professionnelles (RP) et de généralisation de l'évaluation des pratiques dans les établissements de santé en France, le problème de l'appropriation des RP par les professionnels concernés et donc du choix des modalités optimales pour les mettre en œuvre est crucial.

Des travaux ont clairement identifié que les éléments organisationnels et managériaux des établissements de santé, influaient sur la qualité des soins, mesurée par des indicateurs de performance, et plus particulièrement sur l'appropriation des RP. Malgré ces travaux, aucun outil n'avait encore été ni élaboré ni validé en France pour permettre de caractériser le contexte managérial et organisationnel des services d'hospitalisation.

La relation entre ces caractéristiques managériales et organisationnelles et l'appropriation des RP restait donc à explorer de façon plus précise, notamment dans le contexte hospitalier français, avec des outils validés. Par ailleurs, s'il a été démontré que les facteurs individuels influaient sur le degré d'appropriation des RP, on ignorait si ces facteurs pouvaient biaiser la relation entre contexte managérial et organisationnel et appropriation des RP.

L'objectif principal du projet national *Theorem* était d'identifier les éléments managériaux et organisationnels des services d'hospitalisation de médecine polyvalente, et les caractéristiques individuelles des professionnels de santé, associées à l'appropriation des RP.

MATERIEL ET METHODES

Outils

Pour répondre à l'objectif, plusieurs outils quantitatifs distincts ont été élaborés. Le contexte organisationnel et managérial a été identifié au travers de deux questionnaires élaborés par un groupe de travail pluridisciplinaire : un questionnaire de culture organisationnelle (CO) et un questionnaire de faits organisationnels (FO). Un questionnaire individuel interrogeait les professionnels sur des facteurs individuels identifiés au travers d'une revue de la littérature comme associés à l'appropriation des RP.

Au total, 141 variables issues des trois questionnaires étaient potentiellement associées à l'appropriation des RP. La nature des variables est présentée dans le tableau ci-dessous.

Tableau 1- Liste des variables potentiellement explicatives par questionnaire.

Questionnaire	Type	Nombre de variables	Modalités de réponse
Individuel	Profils des professionnels (âge, sexe, formation…) Connaissance et attitudes vis-à-vis des RP Pratiques cliniques et activité perçue Eléments de la personnalité	21 variables	Nominales Echelle d'accord
Culture Organisationnelle (CO)	17 sous-dimensions autour de : 1) Engagement dans le service 2) Résultats perçus du fonctionnement du service 3) Management du service 4) Vie relationnelle et la communication au sein du service 5) Relations avec le patient et sa famille 6) Soutien du chef de service	79 variables	Echelle d'accord à 5 modalités
Faits Organisationnels (FO)	Facteurs structurels Au niveau établissement : Coopération, Activité Equipement, Politique qualité, Veille des RP Au niveau service : Capacité, Spécialité, Activité, Equipement, Ressources humaines, Disponibilités personnels spécialisés	41 variables	Nominales

Le degré d'appropriation des RP a été mesuré via des scores calculés à partir de cas cliniques. Les scores étaient une note comprise entre 0 et 10. Les cas cliniques ont été élaborés par un groupe de praticiens médecins et infirmiers. Ils portaient sur trois types de prise en charge cliniques de patients hospitalisés : la prise en charge de la douleur, la prise en charge de l'insuffisance cardiaque, la prise en charge du diabète. Deux cas cliniques ont été construits pour chaque type de prise en charge et chaque catégorie professionnelle soignante (médecin, infirmier diplômé d'Etat - IDE, aide soignant - AS), constituant ainsi une banque de 18 cas cliniques.

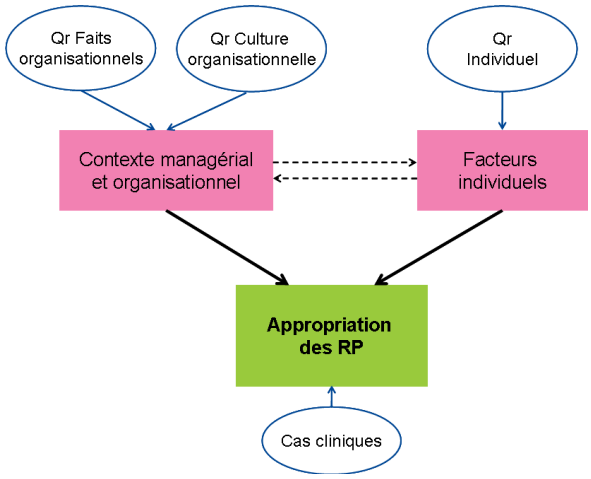


Figure 1 – Schéma de présentation du projet *Théorem*

Recueil de données

Les données ont été recueillies au sein de 36 services de médecine polyvalente d'établissements de santé de tous types (Centres Hospitaliers - CH, Centres Hospitaliers Universitaires - CHU, Cliniques), sélectionnés par tirage au sort, répartis au sein de cinq régions françaises (Aquitaine, Bretagne, Franche-Comté, Poitou-Charentes, Rhône-Alpes). L'échantillon d'étude était constitué de l'ensemble des professionnels ayant une activité de soins, soit les médecins, IDE et AS des services participants. Chaque professionnel a répondu à trois cas cliniques (un par type de prise en charge) attribués aléatoirement, au questionnaire de CO et au questionnaire individuel. Le questionnaire de FO a été complété par un collectif du service. Le recueil a été organisé de façon à garantir la confidentialité totale des données.

Analyses statistiques

Au préalable, des analyses descriptives des différents outils ont été réalisées. Le questionnaire de CO a bénéficié d'une validation métrologique.

L'analyse explicative a été réalisée dans le but de rechercher les facteurs individuels et organisationnels influençant l'appropriation des RP mais aussi d'étudier le sens et le degré d'association de ces facteurs. Une stratégie de modélisation reposant sur des modèles linéaires multivariés à effets aléatoires a été utilisée. Notre variable à expliquer, le degré d'appropriation des RP, était mesurée par trois scores, un pour chaque type de prise en charge des cas cliniques. L'analyse multivariée a permis de prendre en compte conjointement ces trois scores. Les variables des trois questionnaires étaient des facteurs explicatifs potentiels, les uns recueillis au niveau de chaque professionnel et les autres recueillis au niveau du service. Afin de prendre en compte ces données hiérarchiques, un modèle à intercept aléatoire a été utilisé. Le logiciel SAS Proc MIXED permettait de faire un modèle linéaire mixte multivarié.

$$\begin{pmatrix} ScoreC_{ij} \\ ScoreD_{ij} \\ ScoreDl_{ij} \end{pmatrix} = \begin{pmatrix} \beta_0' \\ \beta_0'' \\ \beta_0''' \end{pmatrix} + \begin{pmatrix} \beta_1' \\ \beta_1'' \\ \beta_1''' \end{pmatrix} X_{ij} + \begin{pmatrix} \gamma_{0i}' \\ \gamma_{0i}'' \\ \gamma_{0i}''' \end{pmatrix} + \begin{pmatrix} \varepsilon_{ij}' \\ \varepsilon_{ij}'' \\ \varepsilon_{ij}''' \end{pmatrix} \quad \text{avec i = service = 1...n et j = individu = 1...n}_i ,$$

$$X_{ij} = \text{matrice des variables explicatives} = \begin{bmatrix} X_{ij}' & 0 & 0 \\ 0 & X_{ij}'' & 0 \\ 0 & 0 & X_{ij}''' \end{bmatrix}, \quad \begin{pmatrix} \gamma_{0i}' \\ \gamma_{0i}'' \\ \gamma_{0i}''' \end{pmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\gamma_0'} & \sigma_{\gamma_0'\gamma_0''} & \sigma_{\gamma_0'\gamma_0'''} \\ \sigma_{\gamma_0'\gamma_0''} & \sigma^2_{\gamma_0''} & \sigma_{\gamma_0''\gamma_0'''} \\ \sigma_{\gamma_0'\gamma_0'''} & \sigma_{\gamma_0''\gamma_0'''} & \sigma^2_{\gamma_0'''} \end{bmatrix} \right),$$

$$\begin{pmatrix} \varepsilon_i' \\ \varepsilon_i'' \\ \varepsilon_i''' \end{pmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\varepsilon'} & 0 & 0 \\ 0 & \sigma^2_{\varepsilon''} & 0 \\ 0 & 0 & \sigma^2_{\varepsilon'''} \end{bmatrix} \right) \text{et} \begin{pmatrix} \gamma_{01}' \ldots \gamma_{0n}' \\ \gamma_{01}'' \ldots \gamma_{0n}'' \\ \gamma_{01}''' \ldots \gamma_{0n}''' \end{pmatrix} + \begin{pmatrix} \varepsilon_1' \ldots \varepsilon_n' \\ \varepsilon_1'' \ldots \varepsilon_n'' \\ \varepsilon_1''' \ldots \varepsilon_n''' \end{pmatrix}$$

La stratégie de modélisation était inspirée de celle de Hosmer-Lemeshow avec une méthode ascendante puis descendante.

Dans un premier temps, un modèle « zéro » a été déterminé en cherchant la structure la plus adaptée à nos données. Les variables d'ajustement et de confusion ont été recherchées, le modèle multivarié et les effets aléatoires ont été testés.

Dans un second temps, les variables potentiellement explicatives, ayant moins de 20% de données manquantes et significatives au seuil de 25% dans un modèle multivarié simple, ont été sélectionnées. Au cours de cette étape, un test du rapport de vraisemblance a permis de conclure si l'effet de chaque variable était identique quel que soit le score (estimation globale) ou s'il était différent pour chacun des scores (trois estimations). Les hypothèses suivantes ont été testées :

H0 : $(\beta_1' = \beta_1'' = \beta_1''') = \beta_1^{\circ}$

H1 : au moins un des β_1' ou β_1'' ou β_1''' est différent

Enfin, toutes les variables sélectionnées lors des analyses simples ont été introduites dans un modèle initial multivarié multiple. Les variables non significatives au seuil de 5% ont été éliminées une à une. Concernant les variables du questionnaire de CO, l'ensemble des variables d'une sous-dimension étaient retenues si au moins une des variables était significative. L'adéquation du modèle final a été testée en traçant un diagramme de la répartition des résidus transformés avec la décomposition de Cholesky et un nuage de point entre les valeurs prédites et ces résidus transformés.

RESULTATS

Le modèle « zéro » portait sur 877 professionnels ayant répondus aux cas cliniques. L'effet aléatoire s'est avéré nécessaire. En effet, la différence entre un modèle sans effet aléatoire et un modèle avec effet aléatoire, analysée par un test du rapport de vraisemblance, était significative. Le modèle multivarié était plus adapté que des modèles par score. En effet, les covariances entre les effets aléatoires étaient différentes de zéro. Les analyses précédentes nous ont permis de déterminer que les « numéros de cas clinique » ainsi que la « catégorie professionnelle » étaient des variables d'ajustement. Les effets de ces variables différaient selon les scores, ils ont fait l'objet de trois estimations. Les analyses avec et sans la variable « statut d'établissement » donnaient des estimations très différentes. Elle a été incluse comme variable de confusion avec un effet global sur les trois scores.

Le modèle obtenu en fin de stratégie ascendante comprenait 54 variables sur 141 potentielles et était estimé sur 397 individus. Le modèle final comprenait 17 facteurs identifiés comme en lien avec un meilleur score d'appropriation des RP. Les graphiques sur les résidus ont permis de conclure à une bonne adéquation du modèle.

Au niveau des facteurs individuels, une moyenne d'âge peu élevée et des proportions élevées de professionnels déclarant travailler quotidiennement auprès de patients cardiaques et avoir une expérience en diabétologie favorisaient l'appropriation des RP.

Au niveau des faits organisationnels, les services ayant une grande capacité en nombre de lits, une durée moyenne de séjour modérée (7-10 jours), des prescriptions et résultats d'examens informatisés, un gériatre disponible au moins dans l'établissement, une part élevé de médecins spécialistes par rapport au nombre total de médecins du service et une part moindre d'AS permanentes (non remplaçantes, non intérimaires) par rapport au nombre total d'AS dans le service, avaient une meilleure appropriation des RP.

Enfin, au niveau des caractéristiques de CO, les services favorisant l'appropriation des RP étaient ceux avec des professionnels préoccupés par leur travail au coucher, une gestion de conflits sans recherche à tout prix du consensus, une absence de concurrence entre les médecins, une écoute des paramédicaux par les médecins et un chef de service non stimulant et non mobilisateur.

CONCLUSION

Les modèles mixtes sont désormais couramment utilisés en santé, toutefois l'utilisation de modèles mixtes multivariés est encore rare. Dans le cadre du projet *Theorem*, une modélisation conjointe des trois scores était justifiée. En effet, l'appropriation des RP ne peut pas se mesurer au travers d'un seul type de prise en charge.

Une solution aurait été d'étudier ces trois scores les uns par rapport aux autres, un en temps que variable à expliquer et les deux autres comme variables explicatives. Cependant, il est impossible de faire l'hypothèse qu'un score influence un autre mais que la réciproque n'est pas vraie. Etudier le lien entre nos trois scores sans faire d'hypothèses de sens sur ces liens était donc primordial. Une seconde solution aurait été d'utiliser un modèle d'équations structurelles. Cette méthode fait appel à la notion de variables latentes. Nos scores auraient formés une variable latente. Or, les corrélations entre les scores n'étaient pas très élevées. De plus, les variables potentiellement explicatives des trois questionnaires auraient également dû être transformées en variables latentes. Ceci aurait été possible concernant le questionnaire de CO alors que les questionnaires individuel et de FO ne peuvent pas être traduits en concepts et ne sont donc pas transformables en variable latente.

Ainsi, le modèle linéaire mixte multivarié apparaissait comme une méthode innovante permettant de répondre à des besoins souvent exprimés lors de recherche de facteurs associés à plusieurs variables à expliquer simultanément.

Cette étude, menée sur un large échantillon de services et de professionnels, a permis d'identifier les facteurs les plus fortement associés à l'application des recommandations professionnelles, reflet de la performance clinique des services. Ressort ainsi comme performant un profil de service qui allie un certain dynamisme au travers d'une durée de séjour pas trop importante, un équipement important notamment informatique, une mobilité des aides-soignantes, une présence médicale spécialisée et une vie d'équipe centrée sur la prise en compte de la dimension collective et l'entente. Ces conclusions vont permettre de guider les actions managériales pour optimiser la performance clinique des services au service d'une meilleure prise en charge des patients et du bien être au travail des professionnels.

DANONE RESEARCH
CENTRE DANIEL CARASSO

Application des méthodes bayésiennes à l'analyse d'un événement récurrent

Anissa Elfakir, Jérôme Tanguy, Sébastien Marque

Danone Research, Centre Daniel Carasso, RD128, Avenue de la Vauve, F-91767 Palaiseau Cedex

Mots clés: statistiques bayésiennes, évènements récurrents, données de comptage

RESUME

Afin de démontrer le bénéfice santé d'un produit alimentaire, il est nécessaire d'établir un plan de développement clinique s'appuyant sur un ensemble cohérent d'études pré-cliniques et cliniques. La statistique bayésienne permet d'intégrer dans l'analyse d'une nouvelle étude les connaissances a priori telles que les avis d'experts, ou les résultats des études précédentes. Cette possibilité de capitaliser sur de l'information disponible et hétérogène permet notamment d'accélérer le processus de décision dans le développement clinique [1]. Ces méthodes semblent d'autant plus attrayantes qu'elles bénéficient ces dernières années d'un contexte réglementaire favorable [2] et sont à présent implémentées dans les logiciels couramment utilisés et acceptés par l'industrie.

Dans le cadre d'une étude clinique en nutrition, l'effet d'un produit laitier sur les occurrences multiples d'un évènement a été analysé en utilisant deux familles de méthodes fréquentistes complémentaires (donnée de comptage et donnée de survie) [3-9]. Cette problématique a également été analysée à travers des méthodes bayésiennes afin d'évaluer la sensibilité des résultats.

L'analyse a été faite à l'aide des nouvelles procédures bayésiennes implémentées dans le logiciel SAS® version 9.2 (SAS Institute, Inc, Cary, NC). Une attention particulière à été portée au concept d'élicitation de connaissances expertes. L'influence de la distribution a priori sur le paramètre d'estimation de l'effet produit a été testée en utilisant plusieurs lois informatives et non informatives. Pour évaluer la qualité de l'estimation des paramètres, divers diagnostiques de convergences usuels des chaines de Markov ont été étudiés (graphiques et indicateurs statistiques). Les possibilités d'interprétation offertes par les méthodes bayésiennes et leur l'intérêt par rapport aux méthodes fréquentistes ont été discutées.

DANONE RESEARCH
CENTRE DANIEL CARASSO

Bayesian methods for analysing a recurrent event

Anissa Elfakir, Jérôme Tanguy, Sébastien Marque

Danone Research, Centre Daniel Carasso, RD128, Avenue de la Vauve, F-91767 Palaiseau Cedex

Keywords: Baysiean statistics, recurrent events, count data

SUMMARY

A clinical development plan based on a set of preclinical and clinical studies is needed to prove the health benefit of a food product. Bayesian statistic facilitates the integration of past knowledge, such as expert opinions or results of previous studies, in the analysis of a new study.
The opportunity to capitalise on available and possibly heterogeneous information can accelerate the clinical development and decision making process. Recently, the regulatory environment has looked more favourably on these methods making them even more attractive to implement. This has also been facilitated by the integration of these tools in the most common software's used in the industry today.

For a clinical study in nutrition, the effect of a dairy product on the multiple occurrences of an event was investigated using two frequentist statistical frameworks (count data and survival data) [3-9]. This recurrent issue was also analysed with Bayesian methods as a sensitivity analysis.

The statistical analysis was performed with the new Bayesian procedures implemented in SAS® software release 9.2 (SAS Institute, Inc, Cary, NC). The focus was put on the elicitation of expert knowledge. The impact of the prior distribution on the estimate of the parameter of interest (product effect) was tested using various informative and non informative distributions. The goodness-of-fit of parameter estimates in the model was assessed with different Markov Chain convergence diagnostic tools (plots and statistical indicators). The interpretation of results offered using Bayesian methods was investigated. Finally, the interest of Bayesian methods was discussed and compared to the frequentist methods.

DANONE RESEARCH
CENTRE DANIEL CARASSO

BIBLIOGRAPHIE

1. Spiegelhalter D, Abrams K, Myles J. Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons , 2004.
2. Food and Drug Administration '*Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials'* issued on 5 February 2010'.
3. McCullagh P, Nelder JA. *Generalized Linear Models*, Second Edition. Chapman & Hall: London, 1989.
4. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**(1):1-14.
5. Mullahy J. Specification and Testing of Some Modified Count Data Models, *Journal of Econometrics* 1986; **33**: 341-365.
6. Cook RJ, Lawless JF. *The statistical analysis of recurrent events. Statistics for biology and health*. Springer: New York, 2007.
7. Duchateau L, Janssen P. *The frailty model. Statistics for biology and health*. Springer: New York, 2008.
8. Tanguy J, Elfakir A, Marque S. Strengths and weaknesses of various statistical models to analyse recurrent events. 42èmes Journées de Statistique de la SFdS, Marseille, May 24-28 2010.
9. Tanguy J, Elfakir A, Marque S. Strengths and weaknesses of different statistical models to analyse recurrent events. ISCB Montpellier, France, 29 August-2 September 2010.

Méthodes de correction du degré de signification pour une recherche de codage optimal dans un modèle linéaire généralisé.

Jérémie Riou [a,b] & Benoit Liquet [a]

[a] *Centre de Recherche INSERM U897, Université Bordeaux 2, ISPED, 146 rue Leo Saignat, 33076 Bordeaux Cedex, FRANCE*

[b] *Danone Research, Clinical Studies Plateform, Centre Daniel Carasso, RD 128 - Avenue de la Vauve, 91767 Palaiseau Cedex, FRANCE*

Résumé

En modélisation, la recherche d'un codage optimal pour une variable explicative quantitative entraîne la réalisation de nombreux tests, qui nécessitent de corriger le degré de signification. Nous nous placerons tout d'abord dans le cadre d'une codification binaire pour un modèle linéaire généralisé. Dans ce contexte, il existe une correction exacte du degré de signification du test défini comme le maximum des tests associés aux différents codages. Cette correction sera comparée à la procédure de Bonferroni, d'Efron et aux procédures de rééchantillonnage par permutation et bootstrap. Enfin, nous nous intéresserons à une transformation de la variable en plus de deux classes. Le degré de signification issu du maximum des tests sera ensuite calculé par des méthodes de rééchantillonnage.
Ces méthodes seront comparées par simulations et appliquées sur données réelles. Un package R permettant la réalisation de ces méthodes sera, de plus, proposé.
Mots clefs: procédure de rééchantillonnage, Bonferroni, modèle linéaire généralisé, codage de variable, p-value ajustée, degré de signification, bootstrap, permutation, SAMC

Abstract

In statistical modeling, finding an optimal encoding for a quantitative variable causes the achievement of many tests, which require the correction of significance level. First, we will place in the context of binary coding for a generalized linear model. In this context, it exists an exact correction of test significance level, defined as the maximum of several tests associated with different encodings. This correction will be compared to Bonferroni and Efron correction, and to resampling procedure by permutation and bootstrap. Finally, we generalize this procedure for coding in more than two classes. Significance level of the maximum test will be computed by resampling methods.
These methods will be compare in a simulation study and apply on real data. An R package for the implementation of these methods is also proposed.
Keywords: Resampling procedure, Bonferroni, Generalized Linear Model, Encoding of variable, Adjusted p-value,Significance level, Bootstrap, Permutation, SAMC

1 Introduction

Une pratique courante en modélisation consiste à transformer une variable explicative quantitative en variable catégorielle. Cette transformation se base sur des seuils reconnus scientifiquement. Toutefois, dans de nombreux cas, les seuils ne sont pas connus, il est donc nécessaire de déterminer un codage optimum. Ce choix peut être réalisé en testant de nombreuses combinaisons de seuils jusqu'à obtenir la meilleure d'entre elles. Cette procédure entraîne cependant un problème de multiplicité, nécessitant une correction du degré de signification. Cette gestion de la multiplicité sera étudiée dans le cadre des modèles linéaires généralisés. Pour cela, nous rappelerons tout d'abord la correction exacte existant pour un codage binaire et pour des transformations de type Box-Cox [1]. Puis, nous nous placerons dans le cadre d'un codage en plus de deux classes. Dans ce contexte, il est difficile de déterminer la loi conjointe de tous les tests, et donc de définir une correction exacte. L'idée est donc de comparer des méthodes de rééchantillonnage basées sur le maximum des tests [2] [3] [4], à la correction de Bonferroni. Ces procédures permettent un contrôle du FWER (Family Wise Error Rate) en corrigeant le degré de signification ou en calculant une p-valeur ajustée. Ces méthodes seront disponibles dans un Package R.

2 Contexte statistique

2.1 Le Modèle

Considerons un modèle linéaire généralisé (McCullagh & Nelder, 1989) à p variables explicatives, où les variables réponses $\mathrm{Y}_i (i = 1, \ldots, n)$, sont indépendement distribuées, et possèdent une fonction de densité de famille exponentielle définie comme suit :

$$f_Y(Y_i, \theta_i, \phi) = exp\left\{\frac{\theta_i Y_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)\right\};$$

avec $E[Y_i] = \mu_i = b'(\theta_i)$ et $Var[Y_i] = b''(\theta_i)a(\phi)$.

Le paramètre canonique θ_i est spécifié par :

$$\theta_i = Z_i\gamma + X_i\beta, \tag{1}$$

où X_i est la variable explicative d'intérêt, β le paramètre associé à la variable X_i, $Z_i = (1, Z_i^1, \ldots, Z_i^{p-1})_{p\times 1}$ le vecteur des variables explicatives d'ajustement et $\gamma = (\gamma^0, \gamma^1, \ldots, \gamma^{p-1})_{p\times 1}$ le vecteur des coefficients associés à ces variables.
Nous nous interessons ici à l'effet de la variable X_i, ajustée sur Z_i. Afin de trouver une codification optimale, nous considérons K transformations en m classes de X_i. Le paramètre canonique pour une transformation k s'écrit :

$$\theta_i(k) = Z_i\gamma + X_i^*(k)\beta^k,$$

où $X_i^*(k) = (X_i^1(k), \ldots, X_i^{m-1}(k))_{(m-1)\times 1}$ est défini comme le vecteur des variables indicatrices issue de la transformation catégorielle et $\beta^k = (\beta^1(k), \ldots, \beta^{m-1}(k))_{(m-1)\times 1}$ est le vecteur de coefficient associé à ces indicatrices.

2.2 Procédure basée sur le maximum des tests

Pour chaque transformation, un test sur la nullité du vecteur β_k est réalisé. Nous obtenons, un vecteur des statistiques de tests $T = (T_1, \ldots, T_k)$. Le codage choisi sera celui pour lequel la statistique de test sera la plus grande. Cette procédure consiste à définir une statistique correspondant au maximum des statistiques de tests : $T_{max} = MAX(T_1, \ldots, T_k)$. Pour gérer le problème de multiplicité sous-jacent, nous avons besoin de calculer la p_{valeur} associée à la statistique de test T_{max} définie par :

$$p_{valeur} = P(T_{max} > t_{max}) = 1 - P(T_1 < t_{max}, \ldots, T_{max} < t_{max}).$$

Si l'on connaît la loi conjointe du vecteur T, ainsi que la corrélation entre les tests, il est possible de calculer numériquement cette p_{valeur}. Sinon, on peut approcher cette valeur, par une technique de rééchantillonnage [3], [2].

2.3 Test du score

Pour chaque transformation k, l'hypothèse nulle testée est identique :
$H_0(k) : \beta_1(k) = \beta_2(k) = \ldots = \beta_m(k) = 0$, versus $H_1 : \quad \exists \eta \in \{1, \ldots, m-1\} / \beta_\eta(k) \neq 0$. Pour cela, on décide de réaliser un test du score T_k qui sous l'hypothèse nulle suit asymptotiquement une loi du $\chi^2_{(m-1)\ ddl}$. La statistique de score s'écrit sous la forme:

$$T(k) = U(k)'_{H_0} I(k)^{-1}_{H_0} U(k)_{H_0}; \tag{2}$$

où $U(k)$ et $I(k)$ sont respectivement la fonction de score et la matrice d'information de Fisher sous la contrainte H_0.

3 Méthodes

3.1 Correction exacte du degré de signification pour une transformation binaire [1]

La méthode qui suit est valable pour une transformation dichotomique ainsi que pour une transformation de Box-Cox [5].
Considérons $T(k)$ et $T(l)$ deux tests du score associés aux transformations $X^*(k)$ et $X^*(l)$. Dans le cadre d'une transformation binaire, le test sous H_0 suit asymptotiquement une

loi normale et s'écrit sous la forme :

$$T(k) = \frac{X^*(k)'\hat{R}}{\sqrt{X^*(k)'(I-H)VX^*(k)}};$$

où $\hat{R}$ est le vecteur des résidus estimés sous H_0 dont les composantes sont $\hat{R}_i = Y_i - \hat{\mu}_i$, V est la matrice diagonale définie par $v_{ii} = Var(Y_i)$, et $H = VZ(Z'VZ)Z'$ avec Z la matrice de dimension $n \times p$ définie avec Z_i en ligne.
Nous savons que la loi conjointe des statistiques de tests suit asymptotiquement une loi multivariée normale. Liquet et al. [1], ont défini la corrélation existante entre ces statistiques de test. Il est donc possible de calculer $P(T_1 < t_{max}, \ldots, T_k < t_{max})$ et le degré de signification.

3.2 Procédures pour un codage en plus de deux classes

Dans notre contexte, la statistique du score définie dans l'équation 2, suit asymptotiquement une loi du $\chi^2_{(m-1)\ ddl}$. Pour calculer de manière exacte la statistique du maximum des tests, il est nécessaire de connaître la fonction de distribution de la loi du χ^2 multivariée. Plusieurs méthodes d'approximation de cette fonction existent, mais elles demandent des hypothèses qui ne correspondent pas à notre contexte [6] [7]. Nous avons donc choisi d'utiliser des procédures de rééchantillonnage qui ne nécessitent pas de connaître la fonction de distribution.

3.3 Procédures basées sur le rééchantillonnage

Les procédures de rééchantillonnage ont été développées par Westfall & Young et permettent de contrôler asymptotiquement le FWER. Elles se basent sur le principe du maximum des tests ou du minimum des p-valeurs. Différentes méthodes de rééchantillonnage sous H_0 seront donc mise en place. Il s'agit ici du bootstrap paramétrique, de la permutation, et de l'algorithme SAMC [4]. L'algorithme SAMC a récemment été développé et permet de gagner un temps important pour le rééchantillonnage. Il peut être appliqué dans le cadre de gestion de tests multiples [3].

4 Résultats

Une étude de simulation, ainsi qu'une application sur données réelles ont été réalisées. Elles ont pour but, dans un premier temps, de comparer les méthodes étudiées à la méthode exacte dans le cadre d'un codage binaire, puis de comparer les différentes méthodes dans le cas d'une codification en plus de deux classes. Un package R permettant d'appliquer ces méthodes a été développé.

References

[1] B.Liquet and D.Commenges (2004). Computation of the p-value of the minimum of score tests in the generalized linear model, application to multiple coding. *Statistics & Probability Letters*, 71:33–38.

[2] PH.Westfall and SS.Young (1993). *Resampling-based Multiple Testing*. Wiley.

[3] K.Yu, F.Liang, and J.Ciampa (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics*, 0:1–11.

[4] F.Liang, C.Liu, and RJ.Caroll (2007). Stochastic approximation in monte carlo computation. *Journal of american Statistical Association*, 102:305–320.

[5] VM.Guererro and RA.Johnson. Use the box-cox transformation with binary response models. *Biometrika*, 69:309.

[6] N.Dagupsta and JD.Spurrier (1997). A class of multivariate χ^2 distributions with applications to comparison with a control. *Communication in statistics- Theory and methods*, 26:1559–1573.

[7] T.Royen (1991). Expansions for the multivariate chi-square distribution. *Journal of multivariate analysis*, 38:213–232.

Comparison study of different dose-escalation methods for multiple graded toxicities in oncology

Monia Ezzalfani[1], Marie-Cécile Le Deley[1,2], Sarah Zohar[3]

Institut Gustave-Roussy, Service de Biostatistique et d'Epidémiologie
114 Rue Edouad Vaillant 94805, Villejuif-France
Mail: Monia.ezzalfani@igr.fr; moniezzalfani@yahoo.fr

(1) Institut Gustave Roussy, Biostatistics Department, Vellejuif, France.
(2) Université Paris 11, France.
(3) Inserm, U717, Biostatistics Department, F75010 Paris, France.

Résumé : Le but des essais de phase I est d'identifier la dose à recommander (DR) pour les évaluations ultérieures. Ces essais séquentiels utilisent classiquement un critère binaire de toxicité, ignorant le détail des grades et la multiplicité possible des toxicités. Ce paradigme, développé pour les molécules cytotoxique, est remis en cause par l'émergence des thérapies ciblées, moins toxiques. Le but de notre travail est de comparer les méthodes d'escalade de doses basées sur un score de toxicité, proposées par Yuan(2007), Ivanova(2009) et Chen(2010). Ces méthodes considérant la toxicité comme une variable quasi-continue se différencient par l'inférence statistique et le schéma d'allocation séquentielle des doses: modélisation de la relation dose-score avec inférence Bayésienne pour Yuan; escalade de dose basée sur un algorithme pour Ivanova et Chen. Nous avons évalué les performances des ces méthodes par une étude de simulations sous différents scénarios de relation dose-score de toxicité monotone croissante (150 scénarios tirés selon une loi uniforme). Nous avons approximé la mesure de toxicité par une loi normale tronquée, avec plusieurs hypothèses sur la variance. En moyenne, les résultats de ces méthodes présentent une bonne performance. La méthode de Yuan s'avère sensiblement plus performante en termes de pourcentage de sélection correcte de la DR et d'écart entre la dose sélectionnée et la vraie DR. Les performances des différentes méthodes sont très liées aux hypothèses faites sur la variance. La recherche méthodologique dans ce domaine nécessite une collaboration avec les cliniciens, définissant une mesure de toxicité adéquate, afin de valider la pertinence des méthodes.

Mots clés : Essais cliniques; Les essais de recherche de dose; Phase I; Méthode d'escalade de dose; Toxicité; Thérapies ciblées

abstract: The aim of a phase I oncology trial is to identify the recommended dose (RD) with an acceptable level of toxicity. Dose-Limiting Toxicity (DLT) is the usual endpoint of dose-finding trials, but by only considering toxicity above a predetermined threshold, the use of DLTs ignores much information about lower levels of toxicity which may be relevant for targeted therapies expected to be less toxic. In the present paper, we

choose to highlight and discuss the dose-allocation designs, taking into account multiple toxicities, developed by Yuan et al. (2007), Ivanova and Kim (2009) and Chen et al. (2010). The aim of our work is focused on the comparison of these methods via a large simulation study under multiple relationships between the toxicity measurement and the dose (scenarios). These methods, considering the toxicity as quasi-continuous variable, differ by the statistical inference and the dose-allocation process: the relationship between toxicity scores and doses is modeled by Bayesian approach for the Yuan's method, the dose-escalate is based on an algorithm derived from 'up-and-down' methods for the Ivanaova and Chen's methods. 150 scenarios were randomly generated with uniform distribution to compare between these methods. We choose to approximate the toxicity score distribution by a truncated normal distribution with different variance values. In total the different methods present a good performance. The Yuan's approach seems more performant in terms of the percentage of correct selection of RD and the distance between the estimated dose and the theoretical RD.

1 Introduction

Phase-I dose-finding trials are a key milestone in the development of any new cancer therapy aiming at coming up with safe and efficient drug administration in humans. The aim of phase-I trials is to identify, on a limited number of patients, a safe and reliable recommended dose (RD) for further investigations. For decades of trials assessing cytotoxic agents, the recommended dose was the highest dose associated with an acceptable level of severe toxicity. The emergence of molecularly targeted agents (MTA) in oncology revolutionizes the current phase I paradigm in a variety of ways. In practice, the toxicity response is usually assessed for each body system and graded using a standard grading scale, such as the Common Toxicity Criteria of the National Cancer Institute. A grade of 0 to 4 is assigned reflecting the degree of toxicity for a list of diseases based on the symptoms or physical conditions. It is noticeable that the present version of this established report is 72 pages. Although the toxicity response is intrinsically multidimensional and multiple toxicity events may be observed in a single patient, phase I trial designs are usually based on a binary toxicity endpoint, the Dose-Limiting Toxicity (DLT). Most dose-finding protocols define DLT as a group of grade 3 or 4 non-hematologic and grade 4 hematologic toxicities as well as death (grade 5). This common practice of reducing the toxicity response to a binary indicator is an oversimplification of the complex clinical reality and induces an important loss of information which is particularly detrimental in small size trials: (i) the relative severity of different DLTs is ignored; (ii) a patient experiencing toxic events below the defined threshold of DLT is considered as having no toxicity at all; (iii) multiple toxicities are neglected as only the most severe event is considered to define the DLT. These two latter points are a special matter of concern with targeted agents associated with a different toxicity profile compared to cytotoxic drugs : DLT are

less frequently observed, especially acute myelotoxicity which is foreground with cytotoxic drugs. On the other hand, targeted agents, often called non-cytotoxic agents do not mean free from toxicity .

To define the RD, the dose is typically "escalated" between consecutive cohorts, according to a predefined method of dose-escalation, based on toxicity endpoint. The major methods, used to date, consider the binary DLT data to estimate the RD (up-and-down methods, CRM), however this proposal is inadequate with the emergency of MAT. Some authors have recently proposed different dose-escalation methods based on non binary toxicity outcome. In the present paper, we choose to highlight and discuss the dose-allocation designs developed by Yuan et al. [1], Ivanova and Kim [2] and Chen et al. [3].

2 Methods: Dose-allocation designs

The dose level of a new drug is to be taken from a panel of k discrete dose levels, $D = \{d_1, ..., d_k\}$ (with $k \in K$), with actual toxicity score (TS) assumed to be monotonic and increasing. Let Y_j be a random variable of TS observed at the j^{th} patient, where $Y_j \in [0, \nu]$. Let θ be the target of acceptable toxicity score.

Quasi-CRM approach: Yuan and al. [1] have proposed an extension of the continual reassessment method , called the Quasi-CRM (QCRM), for toxicity scores. The authors have normalized the toxicity scores obtaining values between 0 and 1 and it has been modeled thought a quasi-Bernoulli likelihood. Let Z_{ij} be the normalized score defined as $Z_{ij} = \dfrac{Y_{ij}}{\nu}$ The aim of this method is to find the dose level $d_r \in D$ associated with actual estimated normalized score closest to the normalized target $\tilde{\theta} = \dfrac{\theta}{\nu}$. The quasi-CRM assumes monotonic and increasing relationship between the dose and normalized score. The power function, $\psi(d_i, a) = \alpha_i^a$ was used to model the dose-normalized toxicity score relationship (with α_i defining the normalized working model and $\alpha_i \in [0, 1]$).

Suppose that $g_0(a)$ is the prior distribution for the parameter a. Assume that the last patient is treated at dose level $x_j = d_i$ with normalized score Z_i. The quasi-Bernoulli likelihood is defined as follows:

$$L(x_j, a) = \psi(x_j, a)^{Z_i}(1 - \psi(x_j, a))^{(1-Z_i)}$$

After n patients, the quasi-posterior density for a, having the same form as in the CRM algorithm, is be updated by:

$$g_n(a) = \frac{g_{n-1}(a)L(x_n, a)}{\int_A g_{n-1}(u)L(x_n, u)du}$$

The dose level allocated for the next patient would be the dose associated with a toxicity measurement the nearest to the target ($\tilde{\theta}$). At the end of the trial, the recommended dose, d_r, is the last dose identified by the dose-escalation algorithm.

Unified approach: The unified approach (UA) proposed by Ivanova and Kim [2] can be used for binary, ordinal or continuous endpoints of toxicity. The dose-escalation algorithm was derived from "up-and-down" methods and was based on a t-statistic.

The TS, Y_{ij}, for patient j included at dose d_i, has distribution function $F(\Delta; \mu_i, s_i^2)$, where $(\mu_1, \ldots, \mu_k)$ and $(s_1^2, \ldots, s_k^2)$ are vectors of means and variances $(i = 1, \ldots, k)$. The aim of this method is to identify a dose level $d_r \in D$ such that μ_r closest to θ. Subjects are assigned sequentially starting at the lowest dose level. Let $n(t) = (n_1(t), \ldots, n_k(t))$ $(t \leq n)$ be the number of subjects allocated at each d_i.

Let $T_i(n_i(t))$ denotes the t-statistic, defined as follows:

$$T_i(n_i(t)) = \frac{\bar{Y}_i(n_i(t)) - \theta}{s_i(n_i(t))/\sqrt{n_i(t)}}$$

where $\bar{Y}_i(n_i(t))$ and $s_i(n_i(t))$ are the sample mean and variance of TS computed from all available observations at each d_i, respectively. The dose allocation algorithm is defined as follows:

- if $T_i(n_i(t)) \leq -\Delta$, the next subject is assigned to dose d_{i+1}.
- if $T_i(n_i(t)) \geq \Delta$, the next subject is assigned to dose d_{i-1}.
- if $-\Delta \leq T_i(n_i(t)) \leq \Delta$, the next subject is assigned to dose d_i.

where Δ is the design parameter which is fixed before the beginning of the trial for further details see [2]. Adjustment are made at lower and higher dose levels. At the end of the trial the recommended dose is identified by an isotonic regression.

Dose-allocation method based on Extended Isotonic Design: Chen and al. [3] proposed a design considering a quasi-continuous toxicity and based on an Extended Isotonic Design (EID). Like previous, this method uses a normalized toxicity score Z_{ij}. Let $\hat{q}_i$ be the result of the extended isotonic regression of the toxicity scores at each dose level (see [3] for details) with $i = 1, \ldots, k$. The dose-allocation algorithm is summarized as the follows: The first cohort takes the lowest dose level d_1. After treating a cohort at d_i, the normalized score, Z_{ij}, is calculated and $\hat{q}_i$ is updated of each dose d_i. The dose allocation algorithm is defined as follows:

- If $\hat{q}_k \leq \tilde{\theta}$, then
- if $\tilde{\theta} - \hat{q}_i \leq \hat{q}_{i+1} - \tilde{\theta}$, the next patient is assigned to dose d_{i+1}, where $i \leq K$. Otherwise, the next patient is assigned to dose d_i.
- If $\hat{q}_k \geq \tilde{\theta}$, then
- if $\tilde{\theta} - \hat{q}_{k-1} \leq -\hat{q}_k - \tilde{\theta}$, the next patient is assigned to dose d_{i-1}, where $k \geq 2$. Otherwise, the next patient is assigned to dose d_i.

Adjustment are made at lower and higher dose levels. The steps described as above are iterated until a fixed total number of patients was reached. Doses may not be skipped for the dose escalation or de-escalation. The recommended dose level, d_r is the dose assigned for the last cohort treated.

3 Simulations

we perform 5000 phase I trials, each one on a different dose-toxicity relation (150 scenarios), randomly generated with an uniform distribution. The number of included patients per trial was 25 with a fixed number of dose level of 6 and cohort size of 1. We choose to approximate the toxicity score TS distribution by a truncated normal distribution. Three possible values of variance were compared ($\sigma_{ij}^2 = 1.d_j,\ 1.5.d_j,\ 2.d_j$). For the unified approach (UA) the measure of toxicity varied from 0 to 20. The toxicity target was equal to 4.66 (details are not showen for this result). A sensitivity analysis for the choice of the design parameter Δ showed that $\Delta = 1$ is a reasonable choice (data are not shown in this manuscript). For the Quasi-CRM (QCRM), the prior distribution on the power model for the parameter a is the exponential distribution with mean of 1 and the working model was, obtained by the function "getprior" from the "dfcrm" package in R (http://www.R-project.org) using an empiric model, (0.12, 0.17, 0.23, 0.29, 0.36,0.42).

4 Results

We compared the methods performance in terms of the percentage of the correct selection of the recommended dose (PCS). The PCS is the proportion of trials ending in a correct identification of the recommended dose.Results are summarized by averaging performances over the wide set of situations. Figure 1 shows the variation of the cumulative frequency of the selected RD with the distance between the estimated dose and the theoretical RD (results for variance= $1.d_j$). In total, the different dose-escalation methods have an excellent performance. The Yuan's design seems slightly more performant than other methods. When the estimated dose, is nearly equal to the theoretical RD (Distance from the target value equal to 0), the PCS is around 70% for the Yuan and Ivanova's approach (QCRM and UA) and it is less, around 60% for the Chen's design (EID). The cumulative frequency of the selected RD converge slowly towards 100%; the performance of the methods decreases when the distance between the target value and toxicity score of the estimated target increases.

A sensitivity analysis for the choice of the variance of the TS showed that the performance of the different methods varied with the hypothesis done on the variance (results are available from the authors on simple request).

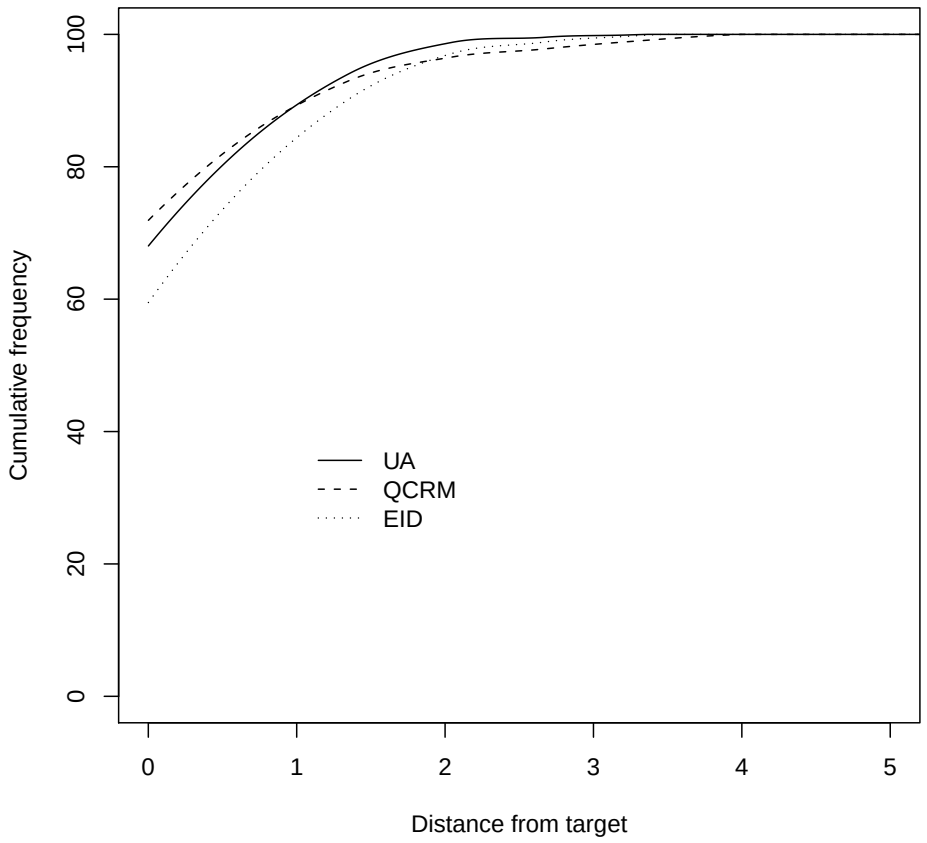


Figure 1: the cumulative frequency of the PCS of the RD according to Yuan, Ivanaova and Chen's methods.

References

[1] Z. Yuan, R. Chappell, and H. Bailey. The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics*, 63:173–179, 2007.

[2] A. Ivanova and S. H. Kim. Dose finding for continuous and ordinal outcomes with a monotone objective function: a unified approach. *Biometrics*, 65:307–315, 2009.

[3] Z. Chen, M. D. Krailo, S. P. Azen, and M. Tighiouart. A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in Phase I clinical trials. *Contemp Clin Trials*, 31(15S (May 20 Supplement)):473–482, 2010.

Fitting an *augmented* Bayesian network to improve a complex quantitative microbial risk assessment model from durability studies.

Sophie Ancelet (a)(b) & Clémence Rigaux (c) & Frédéric Carlin (d)(e) & Christophe Nguyen-thé (d)(e) & Isabelle Albert (c)

(a) AgroParisTech, UMR 518, Equipe MORSE, Paris
(b) EDF R&D, Département Management des Risques Industriels, Chatou
(c) INRA, UR 1204, Met@risk, Paris
(d) INRA, UMR 408, Sécurité et Qualité des Produits d'Origine Végétale, Avignon
(e) Université d'Avignon et des Pays de Vaucluse, UMR 408, Avignon

Résumé:
L'approche par simulations Monte-Carlo est classiquement utilisée en analyse quantitative des risques alimentaires. Lorsque des données sont disponibles, cette approche ne permet pas de "remonter" un réseau bayésien d'Appréciation Quantitative des Risques Microbiologiques (QMRA) pour tirer parti des liens de dépendances entre variables et apprendre efficacement sur les dynamiques d'évolution d'un pathogène alimentaire dans un aliment. Dans un tel problème inverse, construire un réseau bayésien *augmenté* en spécifiant un modèle d'observations puis adopter une démarche inférentielle bayésienne peut s'avérer plus fructueux. Nous montrons qu'il est possible et pertinent de mettre en oeuvre cette approche à partir (i) d'un réseau bayésien nettement plus complexe décrivant la transmission, de la "fourche à la fourchette", d'une population hétérogène de *Bacillus cereus* dans une purée de courgettes (ii) de données d'autocontrôles de vieillissement non directement associées à une variable du QMRA modèle.

Mots-clés : Algorithme MCMC, Appréciation Quantitative des Risques microbiologiques, Données de vieillissement, Incertitude, Inférence bayésienne, Problème inverse, Réseaux bayésiens, Sécurité alimentaire, Simulations Monte-Carlo

Abstract:
The Monte-Carlo simulation approach is traditionally used in food safety quantitative risk assessment. When some data are available, such an approach does not allow "back-calculation" in a Quantitative Microbial Risk Assessment (QMRA) Bayesian Network (BN) to borrow strength across the dependency relationships between the variables and then, efficiently learn about the microbial dynamics of a food-borne pathogen. In such an inverse problem, building an *augmented* BN from an observational model then performing a Bayesian inference may be more fruitful. We aim to show that it is possible and relevant to infer an *augmented* BN in the context of (i) a much more complex QMRA model that describes the dynamics from "farm-to-fork' of a heterogeneous population of *Bacillus cereus* in a courgette purée (ii) additional raw data coming from durability studies and that are not directly linked to one given variable of the QMRA model.

Keywords : Bayesian inference, Bayesian Networks (BN), durability studies, Food safety, Inverse problem, MC simulations, MCMC algorithm, Quantitative Microbial Risk Assessment (QMRA), Uncertainty

1 Introduction

L'approche par simulations Monte-Carlo (MC) est classiquement utilisée pour étudier le comportement sous incertitude de systèmes complexes, souvent décrits à l'aide de réseaux bayésiens. Cela est notamment le cas en Appréciation Quantitative des Risques (AQR) alimentaires où des scénarios d'évolution de la "fourche à la fourchette" du risque microbiologique, associé à un pathogène alimentaire, sont souvent simulés selon un modèle stochastique dit d'Appréciation Quantitative des Risques Microbiologiques (QMRA). Un tel modèle est en fait un réseau bayésien dont la variabilité/incertitude associée à chaque paramètre d'entrée est quantifiée marginalement à l'aide d'une loi de probabilité. L'approche par simulations MC a pour principale limite d'être unidirectionnelle: elle ne permet pas de "remonter" un réseau bayésien afin de tirer parti des liens de dépendances éventuelles entre variables. Elle ne permet pas d'injecter l'information issues de données brutes associées à des variables aléatoires intermédiaires du réseau car ces variables possèdent déjà une distribution de probabilité induite. Enfin, lorsque le modélisateur utilise des données brutes pour spécifier la loi marginale associée à un paramètre d'entrée, la portée de l'information disponible se limite au seul module directement influencé par ce paramètre.

Lorsque des données renseignent sur certaines variables du QMRA modèle, une démarche inférentielle peut s'avérer plus fructueuse pour comprendre et reconstruire des dynamiques plausibles d'évolution du risque associé à un pathogène alimentaire (Smid et al. (2010)). Pour résoudre ce problème inverse, le réseau bayésien d'origine doit préalablement être *augmenté* avec un modèle d'observations. Le cadre bayésien offre une voie d'approche

cohérente et souvent privilégiée pour inférer de tels réseaux bayésiens dits *augmentés*.

Jusqu'à présent, la démarche inférentielle bayésienne sur réseaux bayésiens *augmentés* n'a été que très peu envisagée en AQR: elle a uniquement été mise en oeuvre sur des QMRA modèles simples i.e., exclusivement basés sur des variables aléatoires discrètes (Barker et al. (2005)) ou limités à quelques modules (Delignette-Muller et al. (2006)) ou à une version simplifiée d'une chaine alimentaire (Albert et al. (2008)). Partant d'un cas d'étude spécifique, nous montrons qu'il est possible d'inférer un réseau bayésien *augmenté* dans le cas beaucoup plus complexe où (i) le QMRA modèle inclut de multiples variables continues impliquées dans une combinaison complexe de processus microbiologiques déterministes et de modèles stochastiques décrivant les étapes successives d'une chaine alimentaire définie de la "fourche à la fourchette" (ii) les données brutes disponibles ne sont pas directement associées à une variable spécifique du QMRA modèle.

2 Le cas d'étude

Nous nous intéressons aux dynamiques d'évolution de la "fourche à la fourchette" du pathogène *Bacillus cereus* (*B. cereus*) dans une purée de courgettes dont la complexité du procédé de préparation et de conditionnement est représentative de celle associée à la plupart des aliments réfrigérés et conditionnés à durée de conservation prolongée.

Afchain et al. (2008) ont récemment proposés un QMRA modèle stochastique pour décrire ces dynamiques d'évolution, pour chaque variante génétique possible, et ce, depuis la contamination initiale dans un lot de courgettes brutes jusqu'à la consommation d'un paquet de purée de courgettes par le consommateur. Ce réseau bayésien complexe est basé sur la combinaison de douze modules conditionnellement indépendants décrivant chaque étape de la chaine alimentaire. Il compte 132 variables dont 66 paramètres d'entrée, 60 variables intermédiaires et 6 variables de sortie! Les variables intermédiaires définissent les concentrations latentes en *B. cereus* par lot (ou paquet) et par groupe génétique à chaque étape de la chaîne alimentaire. D'un module à l'autre, ces variables sont liées soit par des relations fonctionnelles décrivant l'inactivation par la chaleur, la germination ou la croissance du pathogène *B. cereus* soit des relations stochastiques comme l'étape de partitionnement aléatoire d'un lot de purée en 860 paquets. Chaque module est controlé par des paramètres d'entrée (généralement des profils temps-température ou des paramètres biologiques) dont la variabilité/incertitude associée a été préalablement quantifiée avec des distributions de probabilité marginales.

Partant de ce QMRA modèle, nous cherchons à injecter l'information portée par 64 nouvelles données de concentrations en *B. cereus* i.e., non-utilisées pour contruire le modèle. Chaque concentration a été mesurée dans un paquet de purée de courgettes pasteurisé soumis à des conditions de stockage temps-température spécifiques. Ces données, qualifiées d'autocontrôles de vieillissement, proviennent de deux sources différentes: l'INRA (14 paquets) en collaboration avec l'université d'Avignon et l'usine

(50 paquets) où sont produits les paquets de purée de courgettes d'intérêt. 75% des données sont censurées à gauche i.e., inférieures à la limite de détection (LoD) associée au protocole expérimental de comptage des bactéries. Pour chaque paquet positif (i.e., dont la concentration bactérienne associée est supérieure à LoD), un typage des souches *B. cereus* a été réalisé afin d'associer la concentration bactérienne totale mesurée au (ou aux deux) groupe(s) génétique(s) dominant(s) présent(s) dans le paquet . Seules les souches II, IV et VI sont observées. Les concentrations bactériennes associées aux autres groupes génétiques sont alors supposées inférieures au dixième de la concentration totale observée.

3 Construction et inférence bayésienne du réseau bayésien *augmenté*

Le QMRA modèle stochastique proposé par Afchain et al. (2008) ne permet pas de simuler le devenir de *B. cereus* dans des paquets de purée de courgettes pasteurisés soumis à des autocontrôles de vieillissement i.e., à des profils temps-température potentiellement extrêmes par rapport aux conditions standards de stockage du produit. C'est pourquoi les données de concentration ne peuvent être directement associées à une variable du réseau. Dans ce contexte, nous proposons de rattacher le QMRA modèle au modèle d'observations par l'intermédiaire de variables latentes intermédiaires qui décrivent la croissance bactérienne en *B.cereus* depuis l'étape de pasteurisation, décrite par le QMRA modèle, jusqu'à la fin des étapes d'autocontrôles de vieillissement. Nous considérons le même modèle de croissance bactérienne que celui utilisé dans le QMRA modèle.

Pour chaque groupe génétique s, soit C_{js}^{obs} la concentration (en ufc/g) observée en *B.cereus* dans le paquet j et C_{js}^{A} la concentration (en ufc/g) induite dans le paquet j après stockage selon des conditions expérimentales spécifiques d'autocontrôles de vieillissement. Pour tenir compte des erreurs de modélisation et erreurs expérimentales susceptibles d'induire des écarts entre C_{js}^{obs} et C_{js}^{A}, nous considérons un modèle de mélange à deux composantes:

- Pour les concentrations associées à des paquets contaminés tels que $log(C_{js}^{obs}) > \alpha_j$:

$$log(C_{js}^{obs}) \sim \mathcal{N}(log(C_{js}^{A}), \sigma^2)$$

- Pour les concentrations censurées à gauche:

$$log(C_{js}^{obs}) \sim \mathcal{N}(log(C_{js}^{A}), \sigma^2)C(;\alpha_j)$$

où $\mathcal{N}(,)C(;\alpha_j)$ désigne la loi normale censurée à droite par $\alpha_j = LoD_j$ lorsque aucun groupe génétique n'est isolé dans le paquet j ou par $\alpha_j = \frac{log(C_{js}^{obs})}{10}$ pour une concentration censurée associée à un groupe génétique non dominant. σ^2 est un paramètre de variance

inconnu contrôlant le niveau global d'erreur de modélisation et d'erreur expérimentale. En l'absence de connaissances *a priori* sur cette quantité, nous proposons d'assigner une loi *a priori* uniforme U[0,100], soit peu informative, sur σ.

L'inférence bayésienne du réseau bayésien *augmenté* proposé a été réalisée avec un algorithme Monte-Carlo par Chaînes de Markov.

4 Résultats

Les nouvelles données de concentration en *B. cereus* acquises n'ont pas permis de modifier les connaissances *a priori* associées à 80% des paramètres d'entrée du QMRA modèle. Cela peut s'expliquer soit par un manque de signal dans les données soit par une quantification *a priori* suffisamment précise de l'incertitude/variabilité associée à ces paramètres pour ne pas être remise en cause avec des données. Néanmois, environ 20% de ces lois *a priori* ont été fortement mises à jour avec l'information portée par les données. Ces mises à jours ont été confrontées à l'opinion des microbiologistes. Certaines d'entre elles (e.g. températures cardinales minimales des groupes de souches II et VI) ont été validées et engendrent donc une réduction de l'incertitude *a priori* sur les paramètres d'entrée associés. D'autres semblent plutôt mener à des remises en cause intéressantes et justifiables de certaines hypothèses de modélisation. Ainsi, par exemple, l'augmentation du temps de réduction décimale *a posteriori* pour les souches du groupe VI pourrait indiquer que ce groupe de souches soit plus hétérogène que prévu et/ou moins résistant à la chaleur dans un laboratoire que dans l'environnement et les aliments. De même, la forte concentration vers de faibles valeurs, observée pour les lois *a priori* sur les temps de cuisson et de pasteurisation équivalents à 90°C, peut seulement s'expliquer en partie par une réduction d'incertitude. En effet, cette réduction est trop importante pour ne pas remettre en question l'ajustement préalablement réalisé à partir de données expérimentales ainsi que le modèle de conversion utilisé pour décrire la variabilité/incertitude associée à ces temps équivalents. Enfin, les prévalences et concentrations des groupes de souches II et VI (i.e., principalement observées dans les autocontrôles de vieillissement) ont significativement augmenté *a posteriori* par rapport aux valeurs spécifiées *a priori* et ce, à toutes les étapes de la chaîne. Ces mises à jour étaient prévisibles et peuvent être imputées à un écart relatif important entre les données observées et les sorties induites par le QMRA modèle. Ainsi la prévalence du groupe génétique VI après pasteurisation est seulement de 0.03% selon le QMRA modèle alors qu'elle est présente dans au moins 7.8% des paquets analysés!

5 Discussion

Une démarche inférentielle bayésienne appliquée à un réseau bayésien *augmenté* permet de confronter un modèle QMRA, classiquement utilisé via des simulations MC, à des données observées. Il peut en découler soit (i) une réduction de l'incertitude à priori et

donc a fortiori une amélioration des connaissances disponibles sur certains paramètres importants impliqués dans les dynamiques de transmission d'un pathogène alimentaire dans un aliment (ii) une remise en cause, en vue d'une amélioration future, du QMRA modèle lorsque les mises à jour observées ne sont pas réalistes aux yeux des experts.

Le cas d'étude considéré montre que construire un réseau bayésien *augmenté* permettant de tenir compte simultanément des caractéristiques des données collectées et d'un QMRA modèle donné peut conduire à des compromis en termes de choix de modélisation. Ainsi, nous avons dû faire l'hypothèse selon laquelle les 64 données de concentration disponibles, bien que provenant de paquets de purée de courgettes différents, nous renseignent sur une unique configuration possible de tous les paramètres d'entrée du QMRA modèle. Cela signifie que la distribution de probabilité associée à chaque paramètre d'entrée du réseau reflète l'incertitude associée à cette quantité aléatoire mais ne peut être associée à une variabilité inter-lots quantifiable avant partitionnement. Cette hypothèse est incontournable lorsqu'on cherche à construire et inférer un réseau bayésien *augmenté* à partir du QMRA modèle proposé par Afchain et al. (2008). En effet, ce dernier a pour limite principale de ne pas distinguer variabilité inter-lots et incertitude lors de la spécification des distributions associées aux paramètres d'entrée du réseau. Pour tenter d'apprendre sur la variabilité inter-lots à partir des données disponibles, il faudrait envisager une version hiérarchique du modèle d'Afchain et al. (2008).

6 Remerciements

Ce travail a été financé par l'Agence Nationale de la Recherche dans le cadre des projets Quant'HACCP et Ribenut, liés à l'appréciation quantitative des risques microbiologiques en sécurité alimentaire.

Bibliographie

[1] Albert, I., et al. (2008) Quantitative Risk Assessment from Farm to Fork and Beyond: a global Bayesian Approach Concerning Food-Borne Diseases. *Risk Analysis*, 28, 557-571.
[2] Smid, J.H., et al. (2010) Strengths and weaknesses of Monte Carlo simulation models and Bayesian belief networks in microbial risk assessment. *International Journal of Food Microbiology*, 139, 57-63
[3] Barker, G.C. et al. (2005) Germination and growth from spores: variability and uncertainty in the assessment of food borne hazards. *International Journal of Food Microbiology*, 100, 67-76.
[4] Delignette-Muller, M.L. et al. (2006) Use of Bayesian modelling in risk assessment: application to growth of Listeria monocytogenes and food flora in cold-smoked salmon. *International Journal of Food Microbiology*,106, 195-208.
[5] Afchain, A.L. et al. (2008) Improving quantitative exposure assessment by considering genetic diversity of B. cereus in cooked, pasteurised and chilled foods. *International Journal of Food Microbiology*, 128, 165-173.

Statistique Spatiale et Processus Ponctuels

Processus stationnaires isotropes et leurs mesures aléatoires associées,

Alain Boudou and Sylvie Viguier-Pla

Beaucoup de phénomènes naturels peuvent être modélisés par des processus stationnaires isotropes. De nombreux auteurs se sont penchés sur ces derniers. Mentionnons par exemple les travaux de Yadrenko (1983), Crujeiras et al. (2008), Stein (2005), Adler (1981) et Cucala et al. (2006). On sait qu'à tout processus stationnaire, on peut associer une et une seule mesure aléatoire, dont il est la transformée de Fourier. Ainsi, toute propriété d'un processus exprimée dans le domaine temporel a son équivalent dans le domaine fréquentiel. On va s'intéresser dans cet exposé à la structure de la mesure aléatoire associée à un processus stationnaire, lorsqu'il est isotrope. La propriété d'isotropie permet de décomposer la mesure aléatoire en un produit tensoriel de deux mesures aléatoires, l'une concernant le rayon, l'autre la direction. Nous pouvons ainsi définir des classes de processus isotropes, nous en donnons quelques exemples que nous illustrons par simulation.

Estimation adaptative dans le cadre d'une modélisation d'interaction poissonienne et application à des données génomiques,

Laure Sansonnet

L'objet de cette communication est de présenter une approche statistique pour étudier les distances favorisées ou évitées entre deux processus donnés le long d'un génome, des gènes ou des motifs par exemple, suggérant de possibles interactions à un niveau moléculaire. Pour cela, on introduit naturellement une fonction dite de reproduction h qui permet de quantifier les positions préférentielles des motifs, par exemple, que l'on modélise ici par l'intensité d'un processus de Poisson. On s'intéresse alors à l'estimation de cette fonction h que l'on supposera très localisée. En utilisant les bases d'ondelettes (en pratique, la base de Haar) et les techniques de seuillage, on peut construire un estimateur adaptatif de h qui satisfait une inégalité de type oracle. On analysera ensuite les avantages et les inconvénients (et les améliorations envisagées) de cette approche et on appliquera la procédure d'estimation à l'étude de la dépendance entre les sites promoteurs et les gènes chez E. coli.

Statistique asymptotique de processus auto-excitatifs spatio-temporels,

Larissa Valmy and Jean Vaillant

Nous nous intéressons aux processus ponctuels auto-excitatifs introduits par Ogata en 1998 et discutés par Zhuang et al. (2005). Ce modèle peut être vu comme une extension, dans un premier temps du processus ponctuel présenté par Hawkes en 1971 et, dans un second temps du modèle de type épidémique temporel proposé par Ogata en 1988. Il s'agit du modèle spatio-temporel ETAS (Epidemic Type Aftershock Sequence) intégrant l'aspect spatio-temporel et les marques. Une de ses utilisations est le calcul de risques sismiques dans une région. Nous étudions les propriétés de la log-vraisemblance et des estimateurs de maximum de vraisemblance dans le cadre d'une application relative à la sismologie de l'Arc Antillais. Des données de tremblements de terre mesurées entre 1999 et 2004 sont analysées et l'adéquation au modèle ETAS est testée.

Référence : Zhuang J., Ogata Y. and Vere-Jones D. (2005). Diagnostic analysis of space-time branching processes for earthquakes. Chap. 15 (Pages 275-290) of Case Studies in Spatial Point Process Models, Edited by Baddeley A., Gregori P., Mateu J., Stoica R. and Stoyan D. Springer-Verlag, New York. 320 pages.

EBSpat un package R dédié à la simulation et l'estimation autour des processus ponctuels de Gibbs de type plus proches voisins, *Rémy Drouilhet*

Dans cet exposé, nous commencerons brièvement par la présentation des principaux résultats théoriques sur les mesures de Gibbs stationnaires avec interactions du type plus-proche voisins. Le package `R` (toujours en cours de développement) `EBSpat` sera alors présenté. Son objectif est de mettre à la disposition des utilisateurs des outils de simulation et d'estimation des processus ponctuels de Gibbs de type plus proches voisins.

Processus stationnaires isotropes et leurs mesures aléatoires associées

Alain BOUDOU & Sylvie VIGUIER-PLA

Equipe de Stat. et Proba., Institut de Mathématiques de Toulouse, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France

Abstract

Many natural phenomenons can be modelized by isotropic stationary processes. Such processes have been the interest of numerous authors, as for example Yadrenko (1983), Crujeiras et al. (2008), Stein (2005), Adler (1981) and Cucala et al. (2006).

It is well-known that each stationary process can be associated with one and only one random measure, by the Fourier Transform. Then, any property of a process expressed in the time domain have its pending in the frequential domain. We will, in this presentation, focus our interest to the structure of a random measure associated with such a process, when it is isotropic. Indead, the isotropic property allows the decomposition of the random measure as a tensor product of two random measures, one for the radius, and the other for the direction. We study the properties of this structure.

We then can define classes of isotropic processes. We give an example of family of isotropic processes, and we illustrate it by simulation.

Key-words. Stationary processes, Random measure, Tensor product, Isotropy, Spectral measure

AMS Classification. 60G57, 60G10, 60B15, 60H05

Résumé

Beaucoup de phénomènes naturels peuvent être modélisés par des processus stationnaires isotropes. De nombreux auteurs se sont penchés sur ces derniers. Mentionnons par exemple les travaux de Yadrenko (1983), Crujeiras et al. (2008), Stein (2005), Adler (1981) et Cucala et al. (2006).

On sait qu'à tout processus stationnaire, on peut associer une et une seule mesure aléatoire, dont il est la transformée de Fourier. Ainsi, toute propriété d'un processus exprimée dans le domaine temporel a son équivalent dans le domaine fréquentiel. On va s'intéresser dans cet exposé à la structure d'un mesure aléatoire associée à un processus stationnaire, lorsqu'il est isotrope. En effet, la propriété d'isotropie permet de décomposer la mesure aléatoire en un produit tensoriel de deux mesures aléatoires, l'une concernant le rayon, l'autre la direction. Nous étudions les propriétés d'une telle structure.

Nous pouvons ainsi définir des classes de processus isotropes. Nous donnons un exemple de famille de processus isotropes, et nous l'illustrons par simulation.

1 Prérequis

Une mesure aléatoire (m.a.) définie sur ξ, tribu de parties d'un ensemble E, à valeurs dans H, $\mathbb{C}-$espace de Hilbert séparable, est une mesure vectorielle Z telle que $< ZA, ZB >= 0$ pour tout couple (A, B) d'éléments disjoints de ξ.

On montre alors que l'application $\mu_Z : A \in \xi \mapsto \|ZA\|^2 \in \mathbb{R}^+$ est une mesure bornée que nous appellerons mesure spectrale de Z.

L'intégrale stochastique, relativement à la m.a. Z, peut se définir comme l'unique isométrie de $L^2(E, \xi, \mu_Z)$ sur $H_Z = \overline{\text{vect}}\{ZA; A \in \xi\}$ qui à 1_A, associe ZA, pour tout A de ξ. L'image d'un élément φ par cette application est notée $\int \varphi \mathrm{d}Z$.

Si (E', ξ') est un deuxième espace mesurable et f une application mesurable de E dans E', alors l'application $f(Z) : A' \in \xi' \mapsto Z(f^{-1}(A')) \in H$ est une m.a., appelée image de Z par f. Sa mesure spectrale est l'image $f(\mu_Z)$ de la mesure bornée μ_Z par f. Si φ' est un élément de $L^2(E', \xi', \mu_{f(Z)})$, alors $\varphi' \circ f$ appartient à $L^2(E, \xi, \mu_Z)$ et $\int \varphi' \circ f \mathrm{d}Z = \int \varphi' \circ \mathrm{d}f(Z)$.

Si J est une application linéaire (resp. anti-linéaire) de H dans H', qui conserve la norme, H' étant un second $\mathbb{C}-$espace de Hilbert séparable, alors
a) $J \circ Z$ est une m.a., définie sur ξ à valeurs dans H', de mesure spectrale μ_Z;
b) pour tout φ de $L^2(E, \xi, \mu_Z)$, on a $\int \varphi \mathrm{d}J \circ Z = J(\int \varphi \mathrm{d}Z)$ (resp. $\int \varphi \mathrm{d}J \circ Z = J(\int \overline{\varphi} \mathrm{d}Z)$).

2 Produit tensoriel de deux mesures spectrales

Soit donc Z (resp. Z') une m.a. définie sur ξ (resp. ξ'), tribu de parties de E (resp. E'), à valeurs dans le $\mathbb{C}-$espace de Hilbert séparable H (resp. H').

Lorsque φ est un élément de $L^2_H(E', \xi', \mu_{Z'})$, on peut affirmer que
a) $< \varphi(.), h >$ appartient à $L^2(E', \xi', \mu_{Z'})$ pour tout h de H ;
b) l'application $\int \varphi \otimes \mathrm{d}Z' : h \in H \mapsto \int < \varphi(.), h > \mathrm{d}Z'(.) \in H'$ est un opérateur de Hilbert-Schmidt appelé intégrale tensorielle de φ par rapport à la m.a. Z' ;
c) l'application $\varphi \in L^2_H(E', \xi', \mu_{Z'}) \mapsto \int \varphi \otimes \mathrm{d}Z' \in \sigma_2(H, H')$ est anti-linéaire et conserve la norme.

Lorsque ψ appartient à $L^2(E \times E', \xi \otimes \xi', \mu_Z \otimes \mu_{Z'})$, on peut vérifier que
a) $\psi(., x')$ appartient à $L^2_H(E, \xi, \mu_Z)$, pour tout x' de E';
b) l'application $\widehat{\psi} : x' \in E' \mapsto \int \psi(., x') \mathrm{d}Z(.) \in H$ est un élément de $L^2_H(E', \xi', \mu_{Z'})$;
c) l'application $\psi \in L^2(\mu_Z \otimes \mu_{Z'}) \mapsto \widehat{\psi} \in L^2_H(E', \xi', \mu_{Z'})$ est linéaire et conserve le produit scalaire.

On peut donc considérer l'application $A \in \xi \otimes \xi' \mapsto \int \widehat{1}_A \otimes \mathrm{d}Z' \in \sigma_2$. Compte-tenu de ce qui précède, on peut facilement établir que cette application, notée $Z \otimes Z'$, est une m.a. telle que
a) $Z \otimes Z'(A \times A') = ZA \otimes Z'A'$, pour tout (A, A') de $\xi \times \xi'$;
b) $\mu_{Z \otimes Z'} = \mu_Z \otimes \mu_{Z'}$;

c) $\int \psi \mathrm{d}Z \otimes Z' = \int \widehat{\psi} \otimes \mathrm{d}Z'$, pour tout ψ de $L^2(\mu_{Z \otimes Z'})$.

La m.a. $Z \otimes Z'$ présente une grande analogie avec le produit de deux mesures classiques, tant au niveau de la définition que des propriétés. En particulier, les règles d'intégration par rapport à une telle mesure évoquent le théorème de Fubini.

3 Application aux processus isotropes

Maintenant nous allons examiner comment la m.a. $\mathcal{Z}$ associée à un processus stationnaire isotrope $(X_t)_{t \in \mathbb{R}^k}$, à valeurs dans $L^2(\Omega, \mathcal{A}, P)$, c'est-à-dire l'unique m.a. dont il est la transformée de Fourier, est liée à une m.a. du type $Z \otimes Z'$. Pour cela, considérons les espaces $E = \mathbb{R}^k - \{0\}$ et $B = \{h \in \mathbb{R}^k; \|h\| = 1\}$. Par $\mathcal{B}_E$ (resp. $\mathcal{B}_B$, resp. $\mathcal{B}_{\mathbb{R}_+^*}$) nous désignerons la tribu de Borel de E (resp. B, resp. $\mathbb{R}_+^*$), par λ l'application continue $x \in E \mapsto \frac{x}{\|x\|} \in B$, et par n l'application continue $x \in E \mapsto \|x\| \in \mathbb{R}_+^*$. L'application $D : x \in E \mapsto (\lambda x, nx) \in B \times \mathbb{R}_+^*$ et sa réciproque $T : (u, r) \in B \times \mathbb{R}_+^* \mapsto ru \in E$ réalisent un homéomorphisme entre les espaces topologiques E et $B \times \mathbb{R}_+^*$.

Nous appellerons également $\mathcal{Z}$ l'application $A \in \mathcal{B}_E \mapsto \mathcal{Z}A \in L^2(\mathcal{A})$, qui est en fait une restriction de la m.a. $\mathcal{Z}$ et est également une m.a. (car tout élément de $\mathcal{B}_E$ est, en tant que sous-ensemble de $\mathbb{R}^k$, un élément de la tribu de Borel de $\mathbb{R}^k$).

L'isotropie du processus stationnaire $(X_t)_{t \in \mathbb{R}^k}$ peut se définir de plusieurs façons équivalentes. En particulier, elle peut se définir par l'invariance de la mesure spectrale $\mu_{\mathcal{Z}}$ par rotation. On montre alors que cette dernière est telle que

- $\lambda \mu_{\mathcal{Z}} = \mu_{\lambda \mathcal{Z}} = \mu$, mesure uniformément répartie sur B ;
- $D\mu_{\mathcal{Z}} = \mu \otimes n\mu_{\mathcal{Z}} = \mu \otimes \mu_{n\mathcal{Z}}$.

Comme les applications $\mathcal{I}_1 : \varphi \in L^2(B \times \mathbb{R}_+^*) \mapsto \int \varphi \mathrm{d}D\mathcal{Z} \in H_{D\mathcal{Z}}$ et $\mathcal{I}_2 : \varphi \in L^2(B \times \mathbb{R}_+^*) \mapsto \int \varphi \mathrm{d}\lambda\mathcal{Z} \otimes n\mathcal{Z} \in \sigma_2(H_{\lambda\mathcal{Z}}, H_{n\mathcal{Z}})$ sont des isométries. Il en est de même de $\mathcal{I} = \mathcal{I}_1 \circ \mathcal{I}_2^{-1}$.

Pour tout (A_1, A_2) de $\mathcal{B}_B \times \mathcal{B}_{\mathbb{R}_+^*}$, il vient alors $\mathcal{I}(\lambda\mathcal{Z}A_1 \otimes n\mathcal{Z}A_2) = D(\mathcal{Z})(A_1 \times A_2)$.

D'où les égalités : $D\mathcal{Z} = \mathcal{I} \circ (\lambda\mathcal{Z} \otimes n\mathcal{Z})$ et $\mathcal{Z} = T(\mathcal{I} \circ (\lambda\mathcal{Z} \otimes n\mathcal{Z}))$.

Compte-tenu des règles d'intégration, pour tout t de $\mathbb{R}^k$, il vient $\int \mathrm{e}^{i<.,t>} \mathrm{d}\mathcal{Z} = \int (\mathrm{e}^{i<.,t>} \circ T) \mathrm{d}\lambda\mathcal{Z} \otimes n\mathcal{Z}$.

Les différentes règles d'intégration, par rapport à la m.a. $\lambda\mathcal{Z} \otimes n\mathcal{Z}$, permettent d'obtenir différents types de décomposition de $\int \mathrm{e}^{i<.,t>} \mathrm{d}\mathcal{Z}$, selon que l'on intègre d'abord par rapport à $\lambda\mathcal{Z}$ ou par rapport à $n\mathcal{Z}$.

Ainsi, si pour tout t de $\mathbb{R}^k$ on pose $Y_t = \int \mathrm{e}^{<.,t>} \mathrm{d}\lambda\mathcal{Z}(.)$, et pour tout s de $\mathbb{R}$ on pose $N_s = \int \mathrm{e}^{i.s} \mathrm{d}n\mathcal{Z}(.)$, on obtient $\int \mathrm{e}^{<.,t>} \mathrm{d}\mathcal{Z} = \mathcal{I}(\int Y_{-rt} \otimes \mathrm{d}n\mathcal{Z}(r)) = \mathcal{I}(\int N_{<u,t>} \otimes \mathrm{d}\lambda\mathcal{Z}(u))^*$.

Bien entendu, la m.a. $\lambda\mathcal{Z}$ concerne la direction et la m.a. $n\mathcal{Z}$ le rayon.

Nous sommes alors en mesure de définir différents types de processus stationnaires isotropes et de les simuler.

Mots-clés. Mesures aléatoires, Processus stationnaires, Produits tensoriels, Isotropie, Mesures spectrales

4 Exemple simulé

On considère une suite $(V_n)_{n\in\mathbb{Z}}$ de variables aléatoires (v.a.) réelles centrées indépendantes définies dans l'espace probabilisé $(\Omega, \mathcal{A}, P)$. Dans notre simulation, les $(V_n)_{n\in\{-m,\ldots,m\}}$ seront des réalisations indépendantes d'une v.a. Gaussienne $N(0, \sigma^2 = 0.5)$.

On définit alors le processus aléatoire $(X_n)_{n\in\mathbb{Z}}$ comme suit : pour tout n de $\mathbb{N}^*$, $X_n = V_{2n} + iV_{2n+1}$, $X_{-n} = V_{2n} - iV_{2n+1}$, et $X_0 = \sqrt{2}V_0$.

On peut vérifier que $(X_n)_{n\in\mathbb{Z}}$ est un bruit blanc, et que, pour tout n de $\mathbb{Z}$, $\|X_n\| = 1$, et $X_n = (-1)^n\overline{X_{-n}}$.

Rappelons que, puisque $(X_n)_{n\in\mathbb{Z}}$ est une série stationnaire, il existe une m.a. et une seule, nommée $\mathcal{Z}$, définie sur la tribu de Borel $\mathcal{B}_\Pi$ de $\Pi = [-\pi, \pi[$, telle que $X_n = \int \mathrm{e}^{i.n}\mathrm{d}\mathcal{Z}$, pour tout n de $\mathbb{Z}$.

Soit v l'application $\theta \in \Pi \mapsto (\cos\theta, \sin\theta) \in \mathbb{R}^2$, elle est continue et donc mesurable. Alors $v(\mathcal{Z})$ est une m.a. définie sur $\mathcal{B}_{\mathbb{R}^2}$, et nous pouvons considérer la fonction continue aléatoire stationnaire $X_{t_1,t_2} = (\int \mathrm{e}^{i(.t_1+.t_2)}\mathrm{d}v(\mathcal{Z}))_{(t_1,t_2)\in\mathbb{R}^2}$. Cette fonction est isotrope parce que :

$< X_{t_1,t_2}, X_{0,0} >= \int \mathrm{e}^{i(.t_1+.t_2)}\mathrm{d}\mu_{v(\mathcal{Z})} = \int \mathrm{e}^{i(.t_1+.t_2)} \circ v\mathrm{d}\mu_{\mathcal{Z}} = \int \mathrm{e}^{i(t_1\cos\theta+t_2\sin\theta)}\mathrm{d}\mu_{\mathcal{Z}}(\theta) = \int \mathrm{e}^{i\sqrt{t_1^2+t_2^2}\cos\theta}\mathrm{d}\mu_{\mathcal{Z}}(\theta)$.

De plus, on peut montrer qu'elle est à valeurs réelles.

Examinons une approximation de cette fonction. Posons $X^{k,m}_{t_1,t_2}$ la v.a.

$X^{k,m}_{t_1,t_2} = \sum_{n=-m}^{m}\sum_{q=0}^{k-1} \mathrm{e}^{i(t_1\cos(-\pi+q\frac{2\pi}{k})+t_2\sin(-\pi+q\frac{2\pi}{k}))}t_n^{k,q}X_n$, où

$t_n^{k,q} = \frac{(-1)^n}{n\pi}\sin\frac{n\pi}{k}\mathrm{e}^{-i(2q+1)\frac{n\pi}{k}} =< \mathcal{Z}([-\pi + q\frac{2\pi}{k}, -\pi + (q+1)\frac{2\pi}{k}[), X_n >$, pour tout n de $\mathbb{Z}^*$, et $t_0^{k,q} = \frac{1}{k} =< \mathcal{Z}([-\pi + q\frac{2\pi}{k}, -\pi + (q+1)\frac{2\pi}{k}[), X_0 >$.

On montre alors que $X^{k,m}_{t_1,t_2}$ "tend", quand k et m deviennent infiniment grands, vers X_{t_1,t_2}, grâce à l'inégalité :

$$\|X_{t_1,t_2} - X^{k,m}_{t_1,t_2}\| \leq \|X_{t_1,t_2} - X^{k}_{t_1,t_2}\| + \|X^{k}_{t_1,t_2} - X^{k,m}_{t_1,t_2}\| \leq \sqrt{t_1^2+t_2^2}\frac{2\pi}{k} + \frac{\sqrt{2}k}{\sqrt{m}}.$$

Prenant $\varepsilon > 0$, on peut choisir un entier k tel que $\sqrt{t_1^2+t_2^2}\frac{2\pi}{k} < \frac{\varepsilon}{2}$, et un entier m tel que $\frac{\sqrt{2}k}{\sqrt{m}} < \frac{\varepsilon}{2}$. Alors nous avons:

$\|X_{t_1,t_2} - X^{k,m}_{t_1,t_2}\| < \varepsilon$.

Pour notre simulation, nous considérons plusieurs valeurs de m et k. Avec les inégalités précédentes, une précision $\varepsilon = 0.1$ est atteinte pour une valeur de m plus grande que 128000, et une valeur de k plus grande que 125. Cependant, empiriquement, la convergence vers une représentation stable de $X^{k,m}_{t_1,t_2}$ semble être atteinte pour de plus petites valeurs de m.

La figure 1 nous permet de nous faire une idée du comportement du processus simulé $X^{k,m}_{t_1,t_2}$ pour différentes valeurs de m et de k, dans le sous-ensemble $[-10, 10]^2$ de $\mathbb{R}^2$.

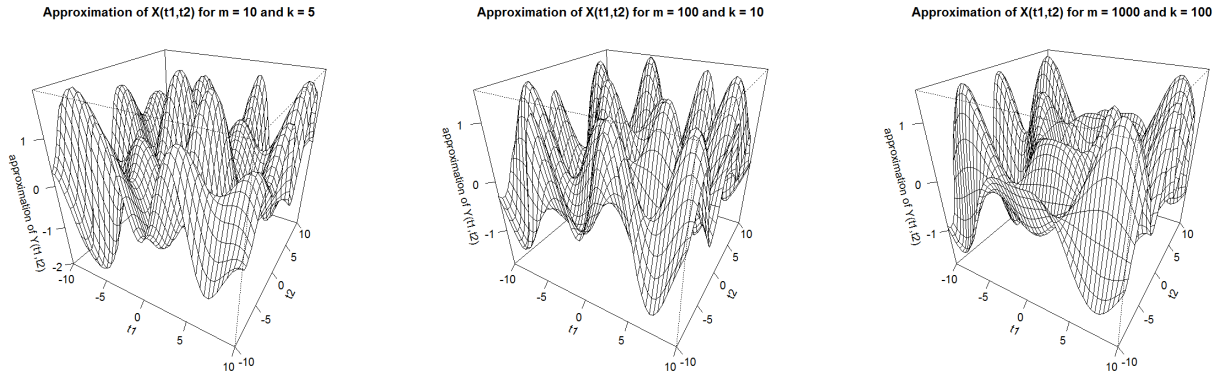


Figure 1 : Processus isotrope approché $X^{k,m}_{t_1,t_2}$ pour différentes valeurs de m et de k

Les graphes de la figure 2 montrent, pour un point arbitraire $(t_1, t_2) = (0.5, 0.5)$, comment les valeurs de $X^{k,m}_{t_1,t_2}$ convergent quand m grandit, k étant fixé, et de même, comment elles convergent quand k grandit, m étant fixé.

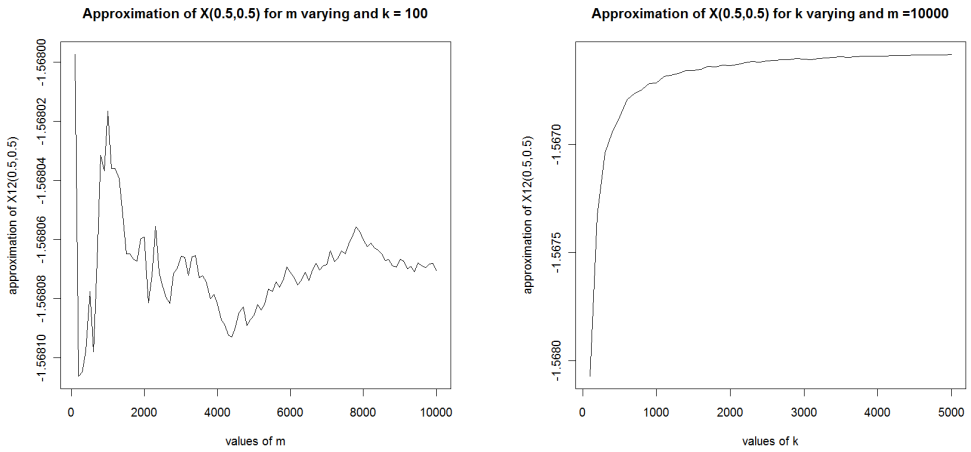


Figure 2 : Convergence du processus isotrope approché en un point de $\mathbb{R}^2$.

La fonction de covariance, qui est uniquement une fonction du rayon, peut aussi être approchée. Plus précisément, $< X_{t_1,t_2}, X_{0,0} >= \int \mathrm{e}^{i\sqrt{t_1^2+t_2^2}\cos\theta} \mathrm{d}\mu_Z(\theta)$ peut être approchée par $\sum_{q=0}^{k-1} \frac{2\pi}{k} \mathrm{e}^{-i\sqrt{t_1^2+t_2^2}\cos\frac{2\pi q}{k}}$, qui reste une fonction du rayon seulement (figure 3).

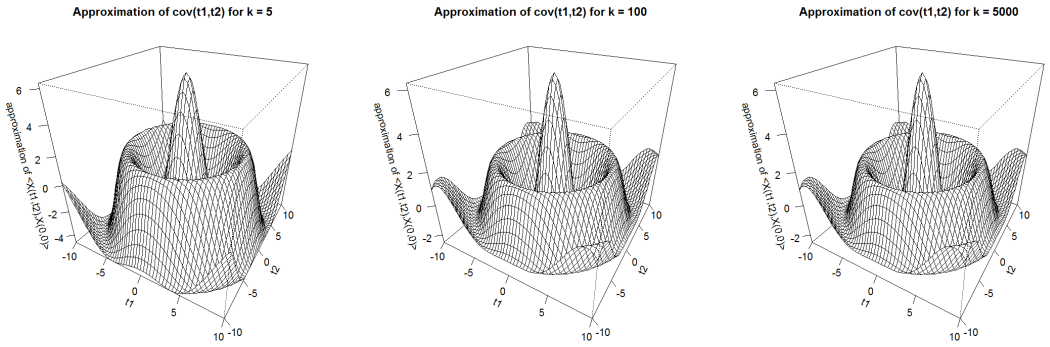


Figure 3 : Fonction de covariance approchée pour différentes valeurs de k

D'autres formes de processus isotropes peuvent ainsi être simulées, en remplaçant les v.a. de loi Gaussienne par n'importe quelle autre v.a., à condition que celle-ci admette un moment d'ordre 2, et que sa variance égale 0.5.

Bibliographie

[1] Adler, R. J. (1981). *The geometry of random fields.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Ltd., Chichester.

[2] Azencott, R. and Dacunha-Castelle, D. (1984). *Séries d'observations irrégulières.* Masson, Paris

[3] Boudou, A. and Romain, Y. (2010). On the integral with respect to the tensor product of two random measures. *J. Multivariate Anal.* **2** 385-394.

[4] Choquet, (1969). *Outils topologiques et métriques de l'Analyse mathématique.* Cours rédigé par Claude Meyer, C.D.U., Paris

[5] Crujeiras, R. M., Fernández-Casal, R., González-Manteiga, W. (2008). An L_2L2-test for comparing spatial spectral densities. *Statist. Probab. Lett.* **78** no. 15, 2543-2551.

[6] Cucala, L., Thomas-Agnan, C. (2006). Spacings-based tests for spatial randomness and coordinate-invariant procedures. *Ann. I.S.U.P.* **50** no. 1-2, 31-45

[7] Dacunha-Castelle, D. and Duflo, M. (1982). *Probabilités et Statistiques.* Masson.

[8] Dehay, D. and Monsan, V. (2007) Discrete Periodic Sampling with Jitter and Almost Periodically Correlated Processes. *Stat. Infer. Stoch. Process.* **10** 223-253.

[9] Letac, G. (1982) *Intégration et probabilités. Analyse de Fourier et analyse spectrale. Exercices.* Masson, Paris.

[10] Matilla, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces.* Cambridge, University press.

[11] Rozanov, Yu. A. (1967). *Stationary Random Processes.* Holden-Day, Inc, San Francisco.

[12] Schaefer, H. H. (1974). *Banach lattices and Positive Operators*. Springer-Verlag, New-York.

[13] Shumway, R. H. and Stoffer, Da. S. (2006). *Stationary Time Series Analysis and its Applications*. Springer, New-York.

[14] Stein, M. L. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100** no. 469, 310-321.

[15] Yadrenko, M. I (1983). *Spectral theory of random fields*. Translated from the Russian. Translation Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York.

Estimation adaptative dans le cadre d'une modélisation d'interaction poissonienne et application à des données génomiques

Laure Sansonnet

Département de Mathématiques
Bâtiment 425
Faculté des Sciences d'Orsay
Université Paris-Sud 11
F-91405 Orsay Cedex

Résumé en français

L'objet de cette communication est de présenter une approche statistique pour étudier les distances favorisées ou évitées entre deux processus donnés le long d'un génome, des gènes ou des motifs par exemple, suggérant de possibles interactions à un niveau moléculaire. Pour cela, on introduit naturellement une fonction dite de reproduction h qui permet de quantifier les positions préférentielles des motifs, par exemple, que l'on modélise ici par l'intensité d'un processus de Poisson. On s'intéresse alors à l'estimation de cette fonction h que l'on supposera très localisée. En utilisant les bases d'ondelettes (en pratique, la base de Haar) et les techniques de seuillage, on peut construire un estimateur adaptatif de h qui satisfait une inégalité de type oracle. On analysera ensuite les avantages et les inconvénients (et les améliorations envisagées) de cette approche et on appliquera la procédure d'estimation à l'étude de la dépendance entre les sites promoteurs et les gènes chez *E. coli*.

Résumé en anglais

The subject of this talk is to present a statistical approach to study favored or avoided distances between two given processes along a genome, genes or motifs for instance, suggesting possible interactions at a molecular level. For this, we naturally introduce a so-called reproduction function h that allows to quantify the favored positions of the motifs, for instance, and which is considered as the intensity of a Poisson process. Our first interest is the estimation of this function h assumed to be localized. Using wavelet bases (in particular, the Haar basis for application) and thresholding, we can define an adaptive estimator of h that satisfies an oracle inequality. We will discuss afterwards this approach's pros and cons (and possible improvements) and we will apply our estimation procedure to study the dependence between gene occurrences along the *E. coli* genome and the occurrences of a motif known to be part of the major promoter sites for this bacterium.

Mots-clés : Estimation adaptative, Inégalité oracle, Ondelettes, Processus de Poisson, Seuillage

Le modèle

On cherche à analyser l'influence entre deux motifs donnés le long d'un génome, un motif étant défini par une séquence de lettres dans l'alphabet `{a,c,g,t}`. Comme les génomes sont longs (environ 1 million de bases) et les motifs courts (de 3 à 20 bases), les occurrences d'un motif peuvent être vues comme des points le long du génome. Et par commodité, on travaille dans un cadre continu donc, on considère les occurrences d'un motif le long du génome comme un processus ponctuel le long d'un intervalle $[0;T]$, où T est la longueur normalisée du génome étudié.

Les points dénommés « parents » :
On observe n points : $U_1,\ldots,U_n$ des variables i.i.d. de loi uniforme sur $[0;T]$, qui décrivent les positions du premier mot d'intérêt sur un brin d'ADN.

Les points dénommés « enfants » :
Chaque U_i donne indépendamment naissance à un processus de Poisson N^i d'intensité h recentrée sur le point parent (voir, par exemple, le livre de Kingman (1993)), i.e. d'intensité la fonction $t \longmapsto h(t-U_i)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, qui modélise les emplacements du deuxième mot d'intérêt, liés à la position U_i du premier mot ; on observe donc le processus agrégé

$$N = \sum_{i=1}^{n} N^i, \quad \text{d'intensité la fonction} \quad t \longmapsto \sum_{i=1}^{n} h(t-U_i). \tag{1}$$

La fonction h, inconnue, est appelée « fonction de reproduction ». h peut être à support infini mais $\int h < \infty$. En pratique, on supposera que h est à support borné dans $[-A;A]$, où $A \in \mathbb{N}$.

À partir de l'observation des U_i et de N, on souhaite **estimer** h.

Dans l'article de Gusto et Schbath (2005), le problème biologique est modélisé par un processus de Hawkes, qui prend en compte l'apparition spontanée (un enfant peut être orphelin) et l'auto-excitation (un enfant peut donner naissance à un autre enfant). Ici, dans le cadre de notre modèle, chaque enfant est issu d'un parent. Mais, dans le contexte biologique, il n'est pas satisfaisant de supposer connu le parent dont est issu chaque enfant. Ce modèle, via la fonction de reproduction h, permet de quantifier les positions préférentielles des enfants par rapport à leur parent.

Résultats généraux

On suppose que la fonction h est dans $\mathbb{L}_1(\mathbb{R})$ et dans $\mathbb{L}_\infty(\mathbb{R})$. On considère sa décomposition dans une base d'ondelettes :

$$h = \sum_{j\geq -1} \sum_{k\in\mathbb{Z}} \beta_{jk}\tilde{\psi}_{jk}, \quad \text{où} \quad \forall j \geq -1, \forall k \in \mathbb{Z}, \beta_{jk} = \int \psi_{jk}(x)h(x)\,dx. \tag{2}$$

En posant

$$\Lambda = \{\lambda = (j,k) : j \geq -1, k \in \mathbb{Z}\},$$

et pour tout $\lambda = (j,k) \in \Lambda$,

$$\varphi_\lambda = \psi_{jk}, \quad \tilde{\varphi_\lambda} = \tilde{\psi}_{jk} \quad \text{et} \quad \beta_\lambda = \beta_{jk},$$

la décomposition (2) peut se réécrire ainsi :

$$h = \sum_{\lambda \in \Lambda} \beta_\lambda \tilde{\varphi_\lambda} \quad \text{avec} \quad \forall \lambda \in \Lambda, \beta_\lambda = \int \varphi_\lambda(x) h(x)\, dx. \tag{3}$$

En pratique, on utilisera la base de Haar (voir, par exemple, le livre de Härdle, Kerkyacharian, Picard et Tsybakov (1998)) qui est définie par son ondelette père : $\phi(x) = \mathbf{1}_{]0;1]}(x)$ et son ondelette mère $\psi(x) = \mathbf{1}_{]1/2;1]}(x) - \mathbf{1}_{[0;1/2]}(x)$.

On définit pour tout λ dans Λ un estimateur <u>sans biais</u> $\hat{\beta}_\lambda$ de $\beta_\lambda = \int \varphi_\lambda(x) h(x)\, dx$:

$$\hat{\beta}_\lambda = \frac{G(\varphi_\lambda)}{n}, \quad \text{avec} \quad G(\varphi_\lambda) = \int_{\mathbb{R}} \sum_{i=1}^{n} \left[\varphi_\lambda(t - U_i) - \frac{n-1}{n} \mathbb{E}_\pi(\varphi_\lambda(t-U)) \right] dN_t, \tag{4}$$

où π est la loi uniforme sur $[0;T]$ et $\mathbb{E}_\pi(\varphi_\lambda(t-U))$ signifie l'espérance de $\varphi_\lambda(t-U)$ où $U \sim \pi$ (indépendant des $U_1, \ldots, U_n$).
On construit ensuite un estimateur $\tilde{\beta}$ de $\beta = (\beta_\lambda)_{\lambda \in \Lambda}$ en considérant une règle de seuillage

$$\tilde{\beta} = \left(\hat{\beta}_\lambda \mathbf{1}_{|\hat{\beta}_\lambda| \geq \eta_\lambda} \mathbf{1}_{\lambda \in \Gamma} \right)_{\lambda \in \Lambda}, \tag{5}$$

où Γ est un sous-ensemble déterministe connu de Λ et $(\eta_\lambda)_{\lambda \in \Lambda}$ une « bonne » famille de seuils <u>aléatoires</u> (leur expression est compliquée et technique).
Cette famille d'estimateurs seuillés nous donne un estimateur adaptatif de h

$$\tilde{h} = \sum_{\lambda \in \Lambda} \tilde{\beta}_\lambda \tilde{\varphi_\lambda} \tag{6}$$

qui satisfait une inégalité de type oracle :

$$\mathbb{E}\left(\|\tilde{h} - h\|_2^2 \right) \leq C_1 \inf_{m \subset \Gamma} \left\{ \sum_{\lambda \notin m} \beta_\lambda^2 + |m| \left[\frac{1}{n} + \frac{n}{T^2} \right] (\log n)^4 \right\} + C_2 \left[\frac{1}{n} + \frac{n}{T^2} \right], \tag{7}$$

où C_1 et C_2 sont des constantes strictement positives dépendant seulement de $\|h\|_1$, $\|h\|_\infty$ et du support de h.
Autrement dit, l'estimateur seuillé de h a un risque pas plus grand que celui de l'oracle, à un terme logarithmique près.

On remarque que lorsque $n = T$, on retrouve le même résultat que dans le cas poissonien simple (voir l'article de Reynaud-Bouret et Rivoirard (2010)).

Application à des données génomiques

On implémente une procédure (qui fait appel à un algorithme de descente) afin de calculer la famille $\tilde{\beta}$ pour reconstruire la fonction d'intensité h.

On souhaite appliquer cette procédure au traitement de données génomiques chez *E. coli*, également utilisées dans l'article de Gusto et Schbath (2005).
On cherche à expliquer l'influence entre les site promoteurs et les gènes chez *E. coli*. On dispose de la position de 4290 gènes (ce seront les enfants du modèle précédemment présenté) et 1036 occurrences du site promoteur `tataat` (ce seront les parents du modèle précédemment présenté), le long du génome (pour les gènes, les positions sont données à la première base codant la séquence). On souhaite détecter les courtes distances favorisées entre `tataat` et gènes.
L'application de la procédure de seuillage permet de mettre en évidence le fait qu'il y a une préférence d'avoir un gène juste après l'occurrence d'un `tataat`, ce qui s'explique biologiquement.

Bibliographie

[1] Gusto, G. and Schbath, S. (2005) FADO : a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Statistical Applications in Genetics and Molecular Biology*, Vol. 4 : Iss. 1, Article 24.
[2] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998) *Wavelets, Approximation and Statistical Applications*, volume 129 of Lecture Notes in Statistics, Springer, New York.
[3] Kingman, J. F. C. (1993) *Poisson processes*, Clarendon Press · Oxford.
[4] Reynaud-Bouret, P. and Rivoirard, V. (2010) Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic Journal of Statistics*, 4, 172–238.

Statistique asymptotique de processus auto-excitatifs spatio-temporels

Larissa VALMY & Jean VAILLANT

Université des Antilles et de la Guyane, LAMIA, EA4540
BP 592 Campus de Fouillole, 97157 Pointe-à-Pitre cédex

Résumé : Nous nous intéressons aux processus ponctuels auto-excitatifs introduits par Ogata (1998) et discutés par Zhuang et al. (2005). Ce modèle peut être vu comme une extension du processus de Hawkes (1971) et du modèle spatio-temporel ETAS d'Ogata (1988). Une de ses utilisations est le calcul de risques sismiques dans une région. Nous étudions les propriétés de la log-vraisemblance et des estimateurs de maximum de vraisemblance dans le cadre d'une application relative à la sismicité de l'Arc Antillais. Ceci est illustré à partir de données sismiques de l'arc des Petites Antilles.

Abstract : We consider self-exciting point processes introduced by Ogata (1998) and discussed in Zhuang et al. (2005). They are extension of Hawkes processes (1971) and ETAS spatio-temporal models (Ogata, 1988) and are useful for calculating seismic risks in a zone. In this paper, we study their log-likelihood and maximum likelihood estimator properties in a spatio-temporal case study of West Indies sismicity. Seismic data from the Caribbean arc are processed.

Mots-clés : *processus de Hawkes, processus ponctuels spatio-temporels, modèle ETAS, log-vraisemblance, estimateur maximum de vraisemblance*

1 Introduction

Une réalisation d'un processus ponctuel spatio-temporel est un ensemble aléatoire de points, où chaque point représente la date et la localisation d'un évènement (Karr, 1991). Les processus ponctuels sont utilisés depuis des années dans divers domaines tels l'épidémiologie, la neurophysiologie, la sismologie puisqu'ils permettent de modéliser des séries d'occurrences d'évènements (Holden et al., 2003 et Kagan, 2004). Sous certaines conditions, ils sont caractérisés de façon unique par leur processus intensité conditionnelle $\lambda(.)$ (Daley et Vere-Jones, 1988). En particulier, les processus de Hawkes sont auto-excitatifs et leur version temporelle a fait l'objet de nombreuses publications depuis l'article pionnier de Hawkes (1971). Ils prennent en compte l'histoire des évènements antérieurs et, dans les versions spatio-temporelles marquées et ETAS (Ogata, 1988), des coordonnées spatiales, temporelles et des marques.

Nous adoptons l'approche suivante : extension du processus ponctuel de Hawkes temporel au spatio-temporel en tenant compte de l'origine du temps de façon analogue à l'approche temporelle de Puri et Tuan (1986). Nous étudions donc les propriétés asymptotiques des processus spatio-temporels de Hawkes au vu de l'étude faite par Puri et Tuan (1986) ainsi que les travaux de Rathbun (1996).

Dans un premier temps, nous faisons quelques rappels sur les processus de Hawkes temporels et nous nous intéresserons à ces processus dans un cadre spatio-temporel. Nous discutons ensuite des propriétés asymptotiques de la log-vraisemblance et des estimateurs de maximum de vraisemblance.

2 Processus spatio-temporel auto-excitatifs

Hawkes (1971) introduit un processus auto-excitatif N d'intensité conditionnelle :

$$\lambda(t) = \mu + \int_{-\infty}^{t} g(t-s)dN(s)$$

où :

- μ est l'intensité d'arrière plan correspondant, par exemple en sismologie, aux secousses principales
- $\int_{-\infty}^{t} g(t-s)dN(s)$ est l'intensité due au passé du processus (les répliques en sismologie).

Il s'agit donc d'un processus purement temporel. Puri et al. (1986) étudie le fait que, dans la pratique, on n'a pas accès à la totalité de l'histoire du processus et que l'on considère donc cette histoire à partir d'une date initiale choisie égale à zéro. Ils regardent l'existence d'éventuels effets sur l'efficacité des techniques de vraisemblance dus à l'approximation de $\lambda(.)$ par

$$\hat{\lambda}(t) = \mu + \int_{0}^{t} g(t-s)dN(s).$$

telle que

$$0 < \int_{0}^{\infty} g(t)dt < 1.$$

Ogata (1988) a développé le modèle temporel ETAS puis il s'est intéressé au spatio-temporel avec marques (Ogata, 2006). Il considère un ensemble de marques $\mathcal{M}$, un espace $X \subset \mathbb{R}^n$, un réel strictement positif T et pose :

$\forall (t, x, M) \in [0, T] \times X \times \mathcal{M}$

$$\lambda^*(t, x, M) = \lambda(t, x | H_{t^-}) \times J(M) \tag{1}$$

où $J(.)$ est la fonction densité de probabilités des magnitudes des évènements ayant $M_c \leq M$ et où M_c est un seuil de magnitude. H_{t^-} est l'histoire du processus N jusqu'à la date t exclue.

$$\lambda(t, x|H_{t^-}) = \mu(x) + \int_0^t \int_X \int_{\mathcal{M}} \kappa(M) \times g(t-s) \times f(x-\epsilon|M) \times dN(s, \epsilon, M) \qquad (2)$$

Dans le membre de droite de (2), interviennent
- $\mu(.)$: fonction intensité d'arrière-plan,
- $\kappa(.)$: nombre attendu d'évènements déclenchés par un séisme de magnitude M,
- $g(.)$: fonction densité de probabilités des dates d'occurrences des répliques,
- $f(.)$: fonction de répartition spatiale conditionnelle des répliques

avec les conditions

$$\int_X \mu(x) dN(x) < +\infty$$

et

$$\int_0^{+\infty} \int_X \int_{\mathcal{M}} \kappa(M) \times g(t-s) \times f(x-\epsilon|M) \times dN(s, \epsilon, M) < 1.$$

3 Propriétés asymptotiques et log-vraisemblance

Soit N un processus spatio-temporel de processus intensité λ_θ, avec $\theta \in \Theta$, Θ étant un ensemble de paramètres. La fonction de log-vraisemblance correspondant à l'intervalle d'observation $[0, T]$ et l'espace mesuré (X, ν) est

$$L_N(\theta) \propto \int_0^T \int_X \big(\log \lambda_\theta(t, x) dN(t, x) - \lambda_\theta(t, x) \nu(dx) dt\big)$$

et peut être approchée par une quasi-log-vraisemblance

$$\widehat{L_N(\theta)} \propto \int_0^T \int_X \big(\log \widehat{\lambda}_\theta(t, x) dN(t, x) - \widehat{\lambda}_\theta(t, x) \nu(dx) dt\big)$$

Cette approximation utilisée à la place de la log-vraisemblance produit un estimateur de maximum de quasi-log-vraisemblance dont le comportement est étudié dans le cadre d'une application relative à la sismicité de l'Arc Antillais.

4 Applications

Des données sismiques de l'arc des Petites Antilles entre 1999 et 2004 sont traitées et une étude comparative est effectuée pour différents modèles.

Bibliographie

[1] Daley, D., and Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.
[2] Hawkes, A.G. (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 :83-90.
[3] Holden, L., Sannan, S. and Bungum, H. (2003) A stochastic marked point process model for earthquakes. *Natural Hazards and Earth System Sciences*, 3 :95-101.
[4] Kagan, Y. Y. (2004) Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America*, 94(4) :1207-1228.
[5] Karr, A. F. (1991) *Point Processes and their statistical inference*, Probability : pure and applied (2), New York.
[6] Ogata, Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of American Statistical Association*, 83(401) :9-27.
[7] Ogata, Y. (1998) Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2) :379-402.
[8] Ogata, Y. and Zhuang, J. (2006) Space-time ETAS models and an improved extension. *Tectonophysics*, 413 :13-23.
[9] Puri, Madan L. and Tuan, Pham D. (1986) Maximum likelihood estimation for stationary point processes. *Proc. Natl. Acad. Sci. USA*, 83 :541-545.
[10] Rathbun, Stephen L. (1996) Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and and Inference*, 51 : 55-74.
[11] Zhuang, J., Chang, C.-P., Ogata, Y. and Chen Y.-I. (2005). A study on the background and clustering seismicity in the Taiwan region by using a point process model. *Journal of Geophysical Research*, 110 : 275-290.

EBSpat un package R dédié à la simulation et l'estimation autour des processus ponctuels de Gibbs de type plus proches voisins

Rémy Drouilhet

LJK, 1251 avenue centrale, BP 47, 38040 Grenoble Cedex 9.

Résumé

Dans cet exposé, nous commencerons brièvement par la présentation des principaux résultats théoriques sur les mesures de Gibbs stationnaires avec interactions du type plus-proche voisins. Le package `R EBSpat` (toujours en cours de développement) sera alors présenté. Son objectif est de mettre à la disposition des utilisateurs des outils de simulation et d'estimation des processus ponctuels de Gibbs de type plus proches voisins.

Abstract

In this talk, we will first present the main results related to the stationary Gibbs measures with nearest-neigbours interactions. The `R` package `EBSpat` (still in development) will then be introduced. Its goal is to provide simulation and estimation tools for Gibbs point processes based on nearest-neighbours interactions.

Mots-clés : Processus ponctuels spatiaux, package `R`.

La classe des processus ponctuels de Gibbs de type plus proches voisins est intéressante car elle permet d'étendre la classe des processus ponctuels de Gibbs superstables introduite par Ruelle dont les interactions entre points sont naturellement basées sur le graphe complet. Il est désormais possible de considérer des interactions entre points basées sur des graphes de type plus proches voisins (graphe de Delaunay, par exemple). De nombreux résultats théoriques portant sur ce type de processus ont été proposés ces dernières années. La première étape a été d'établir l'existence (voir [1,4] pour les princpales contributions) de ces processus ponctuels stationnaires de Gibbs lorsqu'ils sont définis dans $\mathbb{R}^d$ (d correspondant à la dimension). Une série d'articles (dont [2,3]) traitant de l'estimation des processus ponctuels stationnaires (applicables au cadre des processus ponctuels stationnaires de Gibbs de type plus proches voisins) ont donc été ensuite proposés.

En parallèle des premiers développements théoriques sur l'existence de ce type de processus ponctuels de Gibbs, E. Bertin et R. Drouilhet avaient développé un logiciel `C` de simulation de ces processus ponctuels notamment pour les graphes de Delaunay, des k-plus proches voisins et de Gabriel. A la mémoire d'E. Bertin, un package `R`, nommé `EBSpat`, est en cours de développement pour proposer dans le système `R` d'une part les outils de simulation des processus ponctuels de Gibbs de type plus proches voisins et d'autre part les outils d'estimation basés sur la méthode de la pseudo-vraisemblance ainsi que la méthode de Takacs-Fiksel. Ce package est actuellement complémentaire du très complet package `R spatstat` dont le principal contributeur est Adrian Baddeley. A termes, le

développement de EBSpat essaiera de respecter le mieux possible l'esprit de la version 2 du package spatstat (qui sera, semble-t'il, certainement appelé spatstat2).

Afin d'avoir une meilleure idée sur les principales fonctionnalités du package EBSpat, discutons maintenant d'un exemple de simulation et d'estimation pour un processus ponctuel d'interaction de paires de Delaunay. Les quelques lignes de code R suivantes permettent de simuler une réalisation d'un modèle de Strauss sur graphe de Delaunay.

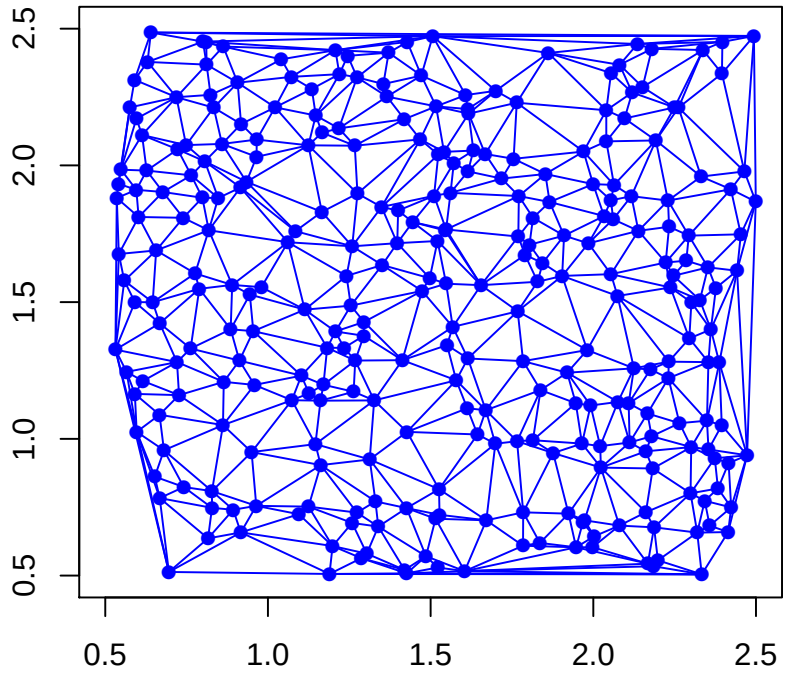
```
> gD <- EBGibbs(~ (-4.61) + Del2(th*(l2<=0.0025),th=0.69),
+              center=c(1.5,1.5),size=2,sizeIn=1.5)
> run(gD) # équivalent ici à "simulate"
```

L'objet gd généré est de la classe EBGibbs. La quantité l2 est prédéfinie et correspond au carré de la longueur d'une arête de Delaunay. Les tailles des domaines intérieur et extérieur sont respectivement fixés à 1.5 et 2. Le potentiel de singleton est fixé à -4.61 et la fonction d'interaction est celle d'un modèle de Strauss basé sur graphe de Delaunay, la portée ayant été fixé à 0.05 et la hauteur du palier à 0.69. Malgré la simplicité de l'exemple choisi ici, nous pouvons remarquer la souplesse offerte à l'utilisateur qui peut librement définir à sa guise la forme de la fonction d'interaction qui sera appliquée à chaque arête de Delaunay. La déclaration générale d'interaction de paires de Delaunay de forme :

$$\sum_{\xi \in Del_2(\varphi)} f(\xi, \theta_1, \cdots, \theta_p) \tag{1}$$

s'effectue en R via l'insertion dans la formule R servant de premier et principal argument de l'instruction EBGibbs, un terme additif de la forme Del2(f(...),theta1=,...,thetaP=) où f(...) est une expression R saisie par l'utilisateur (pour représenter la fonction f dans l'équation (1)) pouvant dépendre en plus des paramètres theta1,...,thetaP (représentant respectivement dans l'équation (1) les paramètres du modèle $\theta_1, \cdots, \theta_p$) des caractéristiques prédéfinies suivantes :

- x: coordonnées des points (ex: x[[1]] et x[[2]][2])
- v: liste des marques (ex: v[[1]]$m et v[[2]]$m2)
- a, da: aires des cellules de Voronoi et ses difference (ex: a[1])
- l, l2: longueur de l'arête de Delaunay et son carré
- ol, ol2: longueur de l'arête duale (orthogonale) de Delaunay et son carré.



Maintenant que nous disposons d'une réalisation du modèle de Strauss pour des arêtes de Delaunay, il nous est possible de proposer les estimations des deux paramètres du modèle correspondant respectivement au potentiel de singleton et à la hauteur du palier (la portée étant supposée ici connue et fixée à la vraie valeur 0.05). Voici les instructions permettant d'obtenir les estimations par la méthode de la pseudo-vraisemblance.

```
> peD <- EBPseudoExpo(gD~Del2(l2<=0.0025),domainSize=1.5)
> run(peD,c(0,0),update=TRUE)
[1] -4.638399 0.8685974
```

La commande `EBPseudoExpo` génère un objet `peD` qui est ensuite appliqué à sa méthode `run` afin d'obtenir les résultats des estimations. La formule établissant la relation entre la réalisation `gD` et le type d'interaction se fait dans l'esprit usuel du système `R`. La syntaxe pour déclarer l'interaction sur le graphe de Delaunay diffère ici de celle utilisée pour la simulation. Comme pour les familles exponentielles, l'interaction du modèle de Strauss se décompose linéairement en les paramètres. La syntaxe à utiliser pour définir les statistiques exhaustives est la suivante : `Del2(f1(...),...,fP(...))` où

`f1(...)`,$\cdots$,`fP(...)` sont les p expressions `R` des statistiques exhaustives dépendant des mêmes caractéristiques (i.e. `x`, `v`, `a`, `da`, `l`, `l2`, `ol` et `ol2`) introduites ci-dessus pour la simulation d'un modèle de paires de Delaunay. Notez aussi que le paramètre de singleton précède toujours les p paramètres associés.

Bibliographie

[1] Bertin, E., Billiot, J.M. et Drouilhet, R. (1999) *Existence of "Nearest-Neighbour" Gibbs Point Models*, Adv. Appl. Prob., 31, 895–909.
[2] Billiot, J.-M., Coeurjolly, J.-F and Drouilhet, R. (2008) *Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes*, Electronic Journal of Statistics, 2 234-264.
[3] Coeurjolly J.-F., Dereudre, D., Drouilhet, R. et Lavancier, F. (2010) *Takacs Fiksel method for stationary marked Gibbs point processes*, Rapport de recherche, HAL, numéro hal-00502004 (soumis et accepté dans Scandinavian Journal of Statistics).
[4] Dereudre, D., Drouilhet, R. et Georgii, H.-0. (2010) *Existence of Gibbsian point processes with geometry-dependent interactions*, arXiv:1003.2875 (soumis et accepté à PTRF).

PLS

PLS et modèle de Cox avec application aux données d'expression de gènes, *Sophie Lambert-Lacroix and Frédérique Letué*

Un aspect important du data-mining dans les données de biopuces est de découvrir la variation moléculaire parmi les cancers. Dans les études de biopuces, le nombre n d'individus est relativement petit comparé au nombre p de gènes par individu (habituellement d'un facteur 100 ou 1000). C'est un challenge important dans le contexte de la prévision des données de survie. Ceci requiert naturellement l'utilisation de procédure de réduction de dimension en même temps que la procédure de prévision.

Dans cet exposé, nous étudions la question de la prévision de la survie dans un contexte de dimension élevée. Nous proposons une nouvelle méthode combinant les méthodes PLS (Partial Least Squares) et la régression de Cox pénalisée de type Ridge, inspirée par Fort et Lambert-Lacroix (2005) dans le contexte de la classification. Nous citons des méthodes existantes basées sur PLS et/ou des techniques de vraisemblances pénalisées, nous soulignons leur intérêt dans certains cas et expliquons leur comportement parfois pauvre sur le plan théorique (en particulier, Bastien (2004) et Li and Gui (2004)). Nous comparons notre procédure avec ces méthodes. Les performances de ces procédures sont illustrées sur deux jeux de données réelles.

A multivariate calibration approach based on support vector regression with direct orthogonal signal correction, *Walid Gani and Mohamed Limam*

Data preprocessing is a fundamental step in calibration for removing noise and unwanted variations arisen from the use of real data. In spectroscopic calibration, direct orthogonal signal correction (DOSC) is a successful preprocessing technique for reducing variants and drifts from high dimensional data. DOSC is usually used with partial least squares (PLS) method. However, the lack of variable selection can spoil the PLS regression. To overcome this issue, we propose the use of support vector regression (SVR) method, which is originally designed to manage large databases. Our approach consists in combining SVR with DOSC for building robust multivariate calibration. The proposed approach is assessed with a real near infrared (NIR) spectroscopic data and compared with the classical DOSC-PLS approach. The results show that DOSC-SVR is better than DOSC-PLS, in terms of root mean square error (RMSE) and $\mathbb{R}^2$.

Sparse PLS deviance residuals, *Philippe Bastien*

The PLS Cox regression has been proposed in the framework of PLS generalized linear regression as an alternative to the Cox model when dealing with highly correlated covariates. However, in high dimensional settings the algorithm becomes computer-intensive and a more efficient algorithm must be used. In this article we propose alternatives both faster and easier to carry out by the direct use of standard procedures which are available in most statistical softwares. Recently, Segal suggested a solution to the Cox-Lasso algorithm when dealing with high dimensional data. Following Segal, we propose a Deviance Residuals based PLS regression (PLSDR) and a Sparse Deviance Residuals based PLS regression (SPLSDR) as an alternative to

the PLS-Cox model in high dimensional settings. The PLSDR and sparse PLSDR algorithms only needs to carry out null deviance residuals using a simple intercept Cox model and use these as outcome in a standard PLS or sparse PLS regression. An application carried out on gene expression from patients with diffuse large B-cell lymphoma shows the practical interest of using deviance residuals as outcomes in PLS or sparse PLS regression when dealing with very many descriptors and censored data.

Analyse en composantes principales et regression PLS quadratiques,

Stéphane Verdun

Dans le cadre de l'Analyse en Composantes Principales (ACP) et de la régression PLS, plusieurs approches ont été proposées pour tenir compte des relations non linéaires existant entre les variables d'un même tableau ou de deux tableaux de données. Parmi ces approches, nous pouvons citer la régression PLS1 quadratique qui a été proposée par Höskuldsson (1992). Dans cette communication, nous étendrons cette démarche au cadre de la régression PLS2 et de l'Analyse en Composantes Principales. Par la suite, nous présentons des applications notamment dans le cadre de la cartographie des préférences.

PLS et modèle de Cox avec application aux données d'expression de gènes

Sophie Lambert-Lacroix [1] & Frédérique Letué [2]

[1] *UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG*
UMR 5525, Grenoble, F-38041, France
[2] *LJK, Université de Grenoble et CNRS, UMR 5224*
51, rue des Mathématiques, B.P. 53, 38041 Grenoble cedex 9, FRANCE

Résumé

Un aspect important du data-mining dans les données de biopuces est de découvrir la variation moléculaire parmi les cancers. Dans les études de biopuces, le nombre n d'individus est relativement petit comparé au nombre p de genes par individu (habituellement d'un facteur 100 ou 1000). C'est un challenge important dans le contexte de la prévision des données de survie. Ceci requiert naturellement l'utilisation de procédure de réduction de dimension en même temps que la procédure de prévision.

Dans cet exposé, nous étudions la question de la prévision de la survie dans un contexte de dimension élevée. Nous proposons une nouvelle méthode combinant les méthodes PLS (Partial Least Squares) et la régression de Cox pénalisée de type Ridge, inspirée par Fort et Lambert-Lacroix (2005) dans le contexte de la classification. Nous citons des méthodes existantes basées sur PLS et/ou des techniques de vraisemblances pénalisées, nous soulignons leur intérêt dans certains cas et expliquons leur comportement parfois pauvre sur le plan théorique (en particulier, Bastien (2004) et Li and Gui (2004)). Nous comparons notre procédure avec ces méthodes. Les performances de ces procédures sont illustrées sur deux jeux de données réelles.

Mots-clés : Analyse des données-data mining, Biostatistique.

Abstract

One important aspect of data-mining of microarray data is to discover the molecular variation among cancers. In microarray studies, the number n of samples is relatively small compared to the number p of genes per sample (usually in thousands). That is a considerable challenge in the context of survival prediction. This naturally calls for the use of a dimension reduction procedure together with the prediction one.

In this talk, the question of survival prediction in such a high dimensional setting is addressed. We propose a new method combining Partial Least Squares (PLS) and Ridge penalized Cox regression, inspired by Fort et Lambert-Lacroix (2005) in

the case of classification. We review the existing methods based on PLS and (or) penalized likelihood techniques, outline their interest in some cases and theoretically explain their sometimes poor behavior (in particular, Bastien (2004) and Li and Gui (2004)). Our procedure is compared with these other methods. The performance of the resulting procedures is illustrated on two real data sets.

Keywords: Data analysis-Data Mining, Biostatistics.

Les technologiques les plus récentes utilisées en biologie et médecine permettent maintenant de recueillir facilement un nombre très important de mesures d'expression de gènes (de l'ordre de plusieurs milliers) d'un individu donné, le nombre d'individu restant lui limité à quelques centaines. Il est donc nécessaire pour les statisticiens de proposer des méthodes de traitement de données efficaces et raisonnables en temps de calcul dans le cas où p le nombre de variables explicatives est très supérieur à n, le nombre d'individus.

Dans un tel contexte, des méthodes comme PLS (Partial Least Squares) ou PCR (Principal Components Regression) ont été classiquement utilisées. Les deux méthodes consistent à trouver des combinaisons linéaires des variables explicatives qui résument l'information contenue dans la totalité des variables (nécessairement redondante). La seconde méthode cherche d'abord ces combinaisons linéaires auxquelles on applique ensuite une régression multiple, alors que la première méthode utilise la variable à expliquer pour choisir les combinaisons linéaires : elle consiste à choisir la combinaison des variables qui maximise la covariance avec la variable à expliquer. Cet algorithme peut s'adapter à la régression pondérée et permet de retrouver l'estimateur du maximum de vraisemblance quand $p < n$. De plus, il fonctionne aussi quand $p >> n$.

Fort et Lambert-Lacroix (2005) ont proposé d'adapter la méthode PLS dans le cas $p >> n$ au modèle de régression logistique où la réponse est binaire. Elles proposent de combiner une étape de Ridge Régression pour régulariser la solution, et une étape PLS pour réduire la dimension. Plus précisement, pour maximiser la vraisemblance, on utilise l'algorithme IRLS, qui revient à effectuer une régression pondérée d'une pseudo-réponse continue, construite à partir des variables explicatives et de la réponse binaire. Mais cet algorithme diverge dans le cas $p >> n$, d'où le recours à une pénalisation ridge. Fort et Lambert-Lacroix proposent donc d'appliquer la méthode PLS à la pseudo-réponse continue, obtenue lors de l'étape RIRLS (pour Ridge Iterated Reweighted Least Squares). Cet algorithme s'avère performant sur plusieurs jeux de données en classification.

Nous nous proposons ici de suivre cette même démarche dans le cadre de données de survie où la réponse est constituée d'une durée éventuellement censurée et d'une indicatrice de censure. Nous nous plaçons dans ce cas dans le cadre du modèle de Cox. Notre algorithme se déroule donc en deux étapes :

1. une étape de régularisation dans laquelle nous maximisons la log-vraisemblance partielle de Cox pénalisée par un critère L_2. Cette étape permet de construire une pseudo-réponse continue et une matrice pondérée.

2. une étape de réduction de dimension dans laquelle nous appliquons la méthode PLS pondérée (avec la matrice de poids trouvée à la première étape) à la pseudo-réponse (trouvée à la première étape).

Une transformation SVD permet de travailler dans un espace de dimension égale au rang de la matrice de départ et non dans en dimension p.

Nous comparons ensuite notre algorithme à ceux trouvés dans la littérature :

- Li and Gui (2004) et Bastien (2004) : nous montrons que ces deux algorithmes sont les mêmes à un choix de poids près.
- Ridge Cox : régularisation sans réduction de dimension.
- PCR Cox : réduction de dimension par PCR, sans régularisation.

La comparaison est faite sur deux jeux de données réelles :

- le jeu de données de cancer du sein de Van't Veer et al. (2002)
- le jeu de données de lymphomes diffus à grandes cellules B (DLBCL) de Rosenwald et al. (2002)

Nous utilisons trois mesures d'évaluation des méthodes :

- la statistique du rapport de vraisemblance
- la variance des résidus matingales
- le score de Brier intégré

Les résultats obtenus montrent que notre algorithme est comparable à celui de Li and Gui (2004) en terme de performance mais bien meilleur en terme de temps de calcul.

Bibliographie

[1] Fort, G. and Lambert-Lacroix, S. (2005) Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7), 1104–1111.

[2] Bastien, P. (2004) PLS-Cox model : application to gene expression. *In COMPSTAT 2004 Proceedings in Computational Statistics*, pages 655–662. Physica, Heidelberg, 2004.

[3] Li, H. and Gui, J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(1), i208–i215.

[4] van Veer, L. J. and Dai, H. and van de Vijver, M. J. and He, Y. D. and Hart, A. A. and Mao, M. and Peterse, H. L. and van der Kooy, K. and Marton, M. J. and Witteveen, A. T. and Schreiber, G. J. and Kerkhoven, R. M. and Roberts, C. and Linsley, P. S. and Bernards, R. and Friend, S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 530–536.

[5] Andreas Rosenwald and George Wright and Wing C. Chan and Joseph M. Connors and Elias Campo and Richard I. Fisher and Randy D. Gascoyne and H. Konrad Muller-Hermelink and Erlend B. Smeland and Jena M. Giltnane and Elaine M. Hurt and Hong Zhao and Lauren Averett and Liming Yang and Wyndham H. Wilson and Elaine S. Jaffe and Richard Simon and Richard D. Klausner and John Powell and Patricia L. Duffey and Dan L. Longo and Timothy C. Greiner and Dennis D. Weisenburger and Warren G. Sanger and Bhavana J. Dave and James C. Lynch and Julie Vose and James O. Armitage and Emilio Montserrat and Armando LÛpez-Guillermo and Thomas M. Grogan and Thomas

P. Miller and Michel LeBlanc and German Ott and Stein Kvaloy and Jan Delabie and Harald Holte and Peter Krajci and Trond Stokke and and Louis M. Staudt, (2002) The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New English Journal of Medicine*, 346, 1937–1947.

A multivariate calibration approach based on support vector regression with direct orthogonal signal correction

Walid Gani & Mohamed Limam

walid.gani@cct.gov.tn & mohamed.limam@isg.rnu.tn

Laboratory of Operational Research, Decision and Control of Processes (LARODEC)
ISG of Tunis, University of Tunis

Abstract

Data preprocessing is a fundamental step in calibration for removing noise and unwanted variations arisen from the use of real data. In spectroscopic calibration, direct orthogonal signal correction (DOSC) is a successful preprocessing technique for reducing variants and drifts from high dimensional data. DOSC is usually used with partial least squares (PLS) method. However, the lack of variable selection can spoil the PLS regression. To overcome this issue, we propose the use of support vector regression (SVR) method, which is originally designed to manage large databases. Our approach consists in combining SVR with DOSC for building robust multivariate calibration. The proposed approach is assessed with a real near infrared (NIR) spectroscopic data and compared with the classical DOSC-PLS approach. The results show that DOSC-SVR is better than DOSC-PLS, in terms of root mean square error (RMSE) and R^2.
Keywords: Data preprocessing ; DOSC ; NIR spectroscopy ; PLS ; RMSE ; SVR.

1 Introduction

NIR spectroscopic data are commonly modeled by multivariate calibration (MVC) procedure, which becomes a standard technique in chemometric. In spectroscopic calibration, data are highly susceptible to noise and unwanted variations. For such reason, preprocessing techniques are applied to remove unwanted systematic variations such as baseline drifts, scatter effects and effects from uncontrolled factors (Gabrielsson and Trygg 2006). Data preprocessing methods are numerous such as normalization methods, smoothing and derivatives (Zeaiter et al. 2005). In recent years, DOSC received a significant attention in chemometric as a powerful tool for data preprocessing (Luypaert et al. 2002 ; Luypaert et al. 2007 ; Zhu et al. 2008). Basically, DOSC is applied with PLS, which is one of the most applied regression technique in spectroscopic calibration (Marx and Eilers, 2002). However, the lack of variable selection can spoil the PLS regression (Höskuldsson 2001). To overcome the shortcoming of PLS, more competitive regression methods have been proposed in MVC such as SVR. In fact, several works have shown that SVR performs

well than PLS (Thissen et al. 2004 ; Hernández et al. 2009 ; Gani et al. 2010). SVR exhibits several advantages in particular its ability to deal with real data and its capacity to model both linear and nonlinear data, thanks to the kernel functions. In addition, SVR was designed to manage high dimensional data, which makes it the most appropriate method to model NIR spectroscopic data.

The paper proposes to combine DOSC with SVR in order to develop a robust MVC. The high computational performance of SVR, due to the solid theoretical background of the method, motivates us to combine DOSC with SVR. The paper is structured as follows. The DOSC method is presented in Section 2. An experimental for assessing the proposed approach is given in Section 3. The results of the experimental are discussed in Section 4. Section 5 summarizes the paper.

2 Direct orthogonal signal correction

The DOSC, developed by Westerhius et al. (2001), is performed as a preprocessing step to improve the MVC model. The method is based on the least squares steps for computing the DOSC components. In the first step of DOSC, the response matrix Y is decomposed into two orthogonal parts: a first part, denoted $\hat{Y}$, representing the projection of Y onto the matrix of predictors X and a second part, denoted E, representing the residual part:

$$Y = P_X Y + A_X Y = \hat{Y} + E, \tag{1}$$

where $P_X = X(X^T X)^{-1} X^T$ projects the column space of X and $A_X = I - P_X$ projects on the orthogonal complement.

In the second step, X is decomposed into two orthogonal parts: a first part having the same range as $\hat{Y}$ and a second part orthogonal to it.

$$X = P_{\hat{Y}} X + A_{\hat{Y}} X. \tag{2}$$

$$A_{\hat{Y}} X = X - P_{\hat{Y}} X. \tag{3}$$

Since E is also orthogonal to X, the columns of $A_{\hat{Y}} X$, span a subspace of X that is orthogonal both to $\hat{Y}$ and $Y = \hat{Y} + E$. After obtaining $A_{\hat{Y}} X$, the principal component t corresponding to the largest singular value, are determined by applying the principal component analysis. The direction of t is expressed as linear combination of X:

$$t = Xr, \tag{4}$$

$$r = X^+ r \tag{5}$$

where X^+ is the Moore-Penrose inverse of X. In order to calculate, the weight vector r, the Moore-Penrose $X^+$is used. This specific inverse is exact meaning that t exactly

equals Xr. Due to this constraint of complete orthogonality, even non-stable directions in X are used to fit the DOSC components t. This leads to an over-fitting. Westerhius et al. (2001), solved this problem by loosening the complete orthogonality constraint. The exact fit of t using the Moore-Penrose inverse X^+ in Equation (5) is loosened by using a generalized inverse X^- which is not completely exact, i.e. $t \approx \tilde{t} = Xr$. The generalized X^- is calculated using a principal component regression (PCR) solution between X and t. The number of principal components for the PCR solution equals the number of singular values of X larger than a tolerance factor, which has to be tuned. Then Equation (5) is modified as:

$$\tilde{r} = X^-, t = X\tilde{r} + e, \tilde{t} = X\tilde{r}. \tag{6}$$

This leads to:

$$X^{DOSC} = X - \tilde{t}\tilde{P}^T = X - X\tilde{r}\tilde{P}^T. \tag{7}$$

$$\tilde{P} = X^T\tilde{t}(\tilde{t}^T\tilde{t})^{-1}. \tag{8}$$

For new samples X_{new} the correction can be performed as follows:

$$X_{new}^{DOSC} = X_{new} - \tilde{r}^T X_{new}\tilde{P}. \tag{9}$$

When DOSC is used, two parameters have to be tuned: the number of DOSC and the tolerance factor, denoted λ. Traditionally, the DOSC is used with a latent variable method such as PLS. However, the fact that PLS model selects a few number of latent variables from the original matrix of predictors, this may disturb the multivariate modeling and reduce the predictive ability of calibration model. Moreover, the number of latent variables should be determined by considering both the curse of dimensionality and loss of data information. Since SVR is a competitive regression method for high dimensional data, therefore it seems to be the most appropriate candidate for a combination with DOSC. Actually, SVR is non-parametric regression method inspired by supervised machine learning. The idea of SVR is based on the computation of a linear function in a high dimensional feature space where the input data are mapped via a nonlinear function. A good review of SVR method is given by Basak et al. (2007).

3 Experimental

In this experimental, we compare the influence of DOSC on two versions of SVR, mainly ϵ-SVR and ν-SVR, and PLS. The NIR spectra of diesel fuels are used in this experimental. These spectra have been measured at Southwest Research Institute on a project sponsored by the U.S. Army. The data are available at *http://www.eigenvector.com*. The training set consists of NIR spectra of 136 diesel fuels. The viscosity of the diesel fuels was obtained

using a separate measurement. The test set consists of 116 diesel fuels. For the SVR models, three types of kernel functions are used:

- The linear kernel is defined as: $K(x_i, x_j) = (x_i.x_j)$.
- The Radial Basis Function (RBF) kernel, which is usually used in the Gaussian form: $K(x_i, x_j) = exp(-\gamma \parallel x_i.x_j \parallel^2)$, where $\gamma = \frac{1}{2\sigma^2}$ and $\sigma > 0$ is a parameter that defines the kernel width.
- The sigmoid kernel, given by: $K(x_i, x_j) = tanh(\kappa\langle x_i.x_j\rangle + \delta)$, where $tanh$ is the hyperbolic tangent function, κ is a parameter controlling the sigmoid shape and δ is a coefficient.

The ϵ-SVR and ν-SVR are implemented in DTREG software and their parameters are optimized using a grid search based on k-fold cross-validation. The optimal latent variables (LV) of PLS are optimized using root mean square error of cross-validation (RMSECV) with leave-one-out procedure. The PLS model is implemented in Minitab software. The DOSC algorithm is implemented in Matlab software. To assess the performance of the regression methods, three performance measures are used

- The RMSE and RMSEP defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \tag{10}$$

  where $\hat{y}$ is the estimated value of y, n the size of samples. The RMSE is termed the root mean square error in calibration (RMSEC) for the calibration set and the root mean square error in prediction (RMSEP) for the test set.

- The coefficient R^2 which is the percentage of the variability of the dependent variable that is explained by the variation of the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

  The R^2 is termed the R^2 of calibration (R_c^2) for the calibration set and the R^2 of prediction (R_p^2) for the test set.

The best regression method corresponds to the lowest value of RMSE and the highest value of R^2.

4 Results and discussion

The DOSC-PLS, DOSC-ϵ-SVR and DOSC-ν-SVR models were built after data correction by DOSC. For the NIR fuel diesel spectra, five DOSC components with a tolerance factor $\lambda = 0.01$ were used for the filtering. The original and the corrected spectra using DOSC are shown in Figure 1.

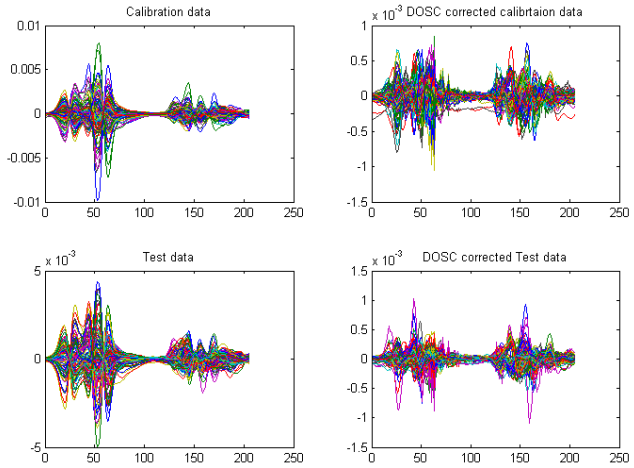


Figure 1: The original and the corrected spectra of NIR fuel diesel data using DOSC.

Table 1: The optimal parameters for ϵ-SVR and ν-SVR.

	ϵ-SVR			ν-SVR		
	linear kernel	RBF kernel	sigmoid kernel	linear kernel	RBF kernel	sigmoid kernel
C	0.3985	451.6005	521.3919	5245.5435	2040.9648	561.7796
ϵ	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
ν	-	-	-	0.0001	0.4629	0.8230
γ	-	0.0033	-	-	0.0033	-
κ	-	-	0.0007	-	-	0.0008
δ	-	-	0.0000	-	-	0.0000
# SV	134	132	133	113	136	136

The DOSC-ϵ-SVR and DOSC-ν-SVR models are performed with three types of kenerl functions and their optimal parameters are given in Table 1. The RBF kernel performs well in comparison with linear and sigmoid kernels. In terms of RMSECV, the optimal

number of LV of DOSC-PLS model is 5, as shown in Figure 2. It is clear from Table 2, that DOSC was successfully applied with SVR method, since DOSC-ϵ-SVR and DOSC-ν-SVR give the lowest values of RMSEC and RMSEP and the highest values of R_c^2 and R_p^2. Before using DOSC method, it is possible to observe that RMSEC and RMSEP values for the two versions of SVR are better than those of PLS. The smallest values of RMSEC resulted from the SVR regression demonstrate that SVR fits well the data and reduces the risk of overfitting. Besides, the smallest values of RMSEP given by SVR show that it has a good power prediction. The ν-SVR outperforms the ϵ-SVR since its R_p^2 which equals 99.3100% is superior to the R_p^2 of ϵ-SVR which equals 96.9610%. After using DOSC, better fit was proven by the DOSC-ν-SVR: higher value of the $R_c^2 = 100\%$ and $R_p^2 = 100\%$. For DOSC-ϵ-SVR model, more reliable values of RMSEC = 0.0027 and RMSEP = 0.0049, are achieved. Even after DOSC filtering, the quality fit of both ϵ-SVR and ν-SVR are better than the quality fit of the PLS method.

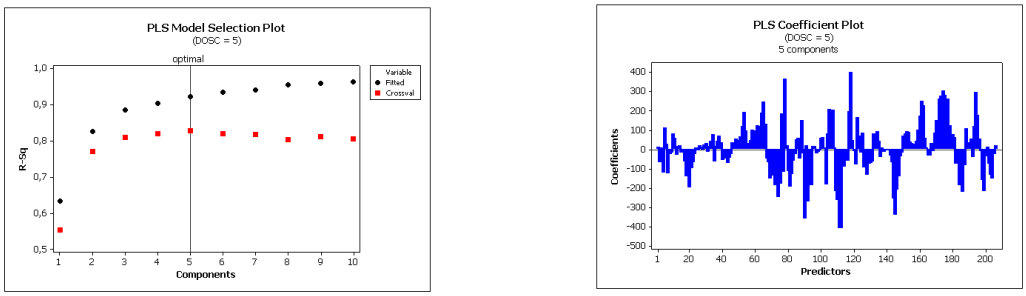


Figure 2: The optimal number of LV and the coefficients of PLS using DOSC.

Table 2: Influence of DOSC on the performance of the regression methods.

	RMSEC	RMSEP	R_c^2	R_p^2
PLS	0.0837	0.1115	96.1948%	91.7873%
ϵ-SVR	0.0701	0.0678	97.3330%	96.9610%
ν-SVR	0.0755	0.0323	96.9090%	99.3100%
DOSC-PLS	0.1200	0.1190	92.1795%	90.6387%
DOSC-ϵ-SVR	0.0027	0.0049	99.9960 %	99.9850%
DOSC-ν-SVR	0.0000	0.0000	100.0000%	100.0000%

5 Conclusion

The paper proposes a strategy which combines DOSC with SVR for building a robust MVC for high dimensional data. The proposed strategy comes to overcome the issues encountered with the classical DOSC-PLS strategy. Over a real experimental, we show that DOSC-SVR outperforms DOSC-PLS in terms of RMSE and R^2. The experimental demonstrated that SVR is easy to implement and more flexible and robust than the PLS method.

Bibliography

[1] Basak, D., Pal, S. and Patranabis, D.C. (2007) Support vecttor regression. *Neural Information Processing - Letters and Reviews*, 11, 203–224.

[2] Gabrielsson, J. and Trygg, J. (2006) Recent developments in multivariate calibration. *Critical Reviews in Analytical Chemistry*, 36, 243–255.

[3] Gani, W., Taleb, H. and Limam, M. (2010) Support vector regression based residual control charts. *Journal of Applied Statistics*, 37, 309–324.

[4] Hernández, N., Talavera, I., Biscay, R.J., Porro, D. and Ferreira, M.M.C. (2009) Support vector regression for functional data in multivariate calibration problems. *Analytica Chimica Acta*, 642, 110–116.

[5] Höskuldsson, A. (2001) Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55, 23–38.

[6] Luypaert, J., Heuerding, S., de Jong, S. and Massart, D.L. (2002) An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of Pharmaceutical and Biomedical Analysis*, 30, 453–466.

[7] Luypaert, J., Heuerding, S., Massart, D.L. and Vander Heyden, Y. (2007) Direct orthogonal signal correction as data pretreatment in the classification of clinical lots of creams from near infrared spectroscopy data.*Analytica Chimica Acta*, 582, 181–189.

[8] Marx, B.D. and Eilers, P.H.C. (2002) Multivariate calibration stability: a comparison of methods. *Journal of Chemometrics*, 16, 129-140.

[9] Thissen, U., Pepers, M., Üstün, B., Melssen, W.J. and Buydens, L.M.C. (2004) Comparing support vector mahines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*, 73, 169–179.

[10] Westerhuis, J.A., Jong, S. and Smidle, A.K. (2001) Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 56, 13–25.

[11] Zeaiter, M., Roger, J.M. and Bellon-Muarel, V. (2005) Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods. *Trends in Analytical Chemistry*, 24, 437–445.

[12] Zhu, D., Ji, B., Meng, C., Shi, B., Tu, Z. and Qing, Z. (2008) The application of direct orthogonal signal correction for linear and non-linear multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 90, 108–115.

SPARSE PLS DEVIANCE RESIDUALS

Philippe Bastien

L'OréaL Research and Innovation, 90 rue du général Roguet, 92583 Clichy cedex
pbastien@rd.loreal.com

The evolution of cancer is more certainly linked to a complex interplay of genes rather than a single gene activity. Multivariate analysis, which can exploit the correlated pattern of gene expression display by genes behaving jointly, such as genes performing the same functions or genes operating along the same pathway, can become a very useful diagnostic tool to determine molecular predictor of survival on the basis of gene expression. However, problems encountered in multiple regressions due to multicollinearity, or ill posed problems with many descriptors and only a few samples, occur in the same way when we are dealing with censored data. The proportional hazard regression model suggested by Cox in 1972 to study the relationship between the time to event and a set of covariates in the presence of censoring is the model most commonly used for the analysis of survival data. However, like multivariate regression, it supposes that there are more observations than variables, complete data, and variables not strongly correlated between them. These constraints are often crippling in practice, as for example in oncology when the expression of several thousands of genes is collected from bio-chips and used as molecular predictors of survival.

Prediction in high-dimensional and low-sample size settings already arose in Chemistry in the eighties. The PLS regression, which could be viewed as a regularization method based on dimension reduction, was developed as a Chemometric tool in an attempt to find reliable predictive models with spectral data. Nowadays, the difficulty encountered with the use of transcriptomic data for classification or prediction, using very large matrices, is of comparable nature. It was thus natural to use PLS regression principles in this new context.

In 1994, Garthwaite showed that PLS regression could be obtained as a succession of simple and multiple linear regressions. In 1999, Tenenhaus proposed a quite similar approach but which can also cope with missing data by using the principle of the NIPALS algorithm. By using this alternative formulation of the PLS regression, Tenenhaus (1999) and Esposito Vinzi and Tenenhaus (2001) have extended the PLS regression to logistic regression by replacing the succession of simple and multiple regressions by a succession of simple and multiple logistic regressions. This algorithm was then generalized by Bastien and Tenenhaus (2001) who extended the PLS regression to any generalized linear regression and to the Cox model as a special case. Further improvements have then been proposed in the case of categorical descriptors with model validation by bootstrap resampling [1]. Applications of the PLS-Cox model have been proposed to predict the occurrence of prematurely greying hair and to predict, based on gene expression, the survival probability of patients with lymphoma.
Since 2001, in order to deal with high dimensional censored survival data, many developments in the framework of the Cox regression have appeared in the literature using regularized approaches. The regularization can be expressed as a dimension reduction, as is the case with the PLS regression by constraining the solution to be located in a subspace named Krylov space giving biased regression coefficients but with lower variance. The regularization can also be expressed by minimization of the negative log partial likelihood subject to some L_1 constraints on the coefficients. This approach has been proposed by Tibshirani in 1997 as an extension of its popular built-in variable selection procedure called least absolute shrinkage and selection operator (Lasso). However the quadratic minimization embedded in the Cox-Lasso procedure cannot be solved in high dimensional settings.

In 2004, Efron et al. proposed a highly efficient procedure called LARS for Least Angle Regression for variable Selection which can be slightly modified to provide solution for the Lasso. Gui and Li, in 2005 , taking advantage of the close connection between LARS and Lasso proposed a solution to the Cox-Lasso procedure in the setting of very high-dimensional data. However, the iterative reweighted least squares strategy used by the LARS-Cox procedure for handling survival endpoints undo much of the computational efficiency of the LARS-Lasso procedure. Recently, Segal has described an estimation strategy that restores the computational efficiency of the LARS-Lasso procedure. He has shown that the expression to be minimized in the Cox-Lasso procedure of Tibshirani can be approached by the deviance residuals sum of squares. He then proposed to perform the original LARS-Lasso algorithm on those specific residuals.

The PLS-Cox algorithm is sequential in the estimation of the weights used in the determination of the PLS components. When the number of descriptors exceeds by far the number of observations, as it is the case with gene expression where the number of genes can reach several tens of thousands, the algorithm becomes computer-intensive and a more efficient algorithm must be used.

Following Segal, an alternative to the PLS-Cox model in high dimensional settings by using deviance residual based PLS regression has been proposed by Bastien [2]. This approach is advantageous both by its simplicity and its efficiency because it only needs to carry out null deviance residuals using a simple intercept Cox model and use these as outcome in a standard PLS regression.

Other recent approaches in the generalised lasso regularization framework are the path approach of Park and Hastie [3] which directly generalized the Lars algorithm to generalized linear models and more recently two gradient based algorithms which does not require a matrix inversion.: the gradient lasso for Cox proportional model of Sohn et al.[4] and the gradient algorithm of Goeman [5]

We propose a new formulation of the PLS Deviance Residual algorithm in the framework of sparse PLS regression by using the Sparse PLS algorithm of Keles et al. [6] based on the deviance residuals of a simple intercept Cox model. Let’s note that in the case of PLS1 the solution becomes the Univariate Soft Thresholding estimate (UST) proposed by Zou and Hastie [6] with their Elastic Net algorithm when the ridge parameter tends to infinity. This is not surprising as the ridge regression tends to one factor PLS regression when the ridge parameter tends to infinity, or if the Elastic Net could be viewed as a combination of LARS and Ridge regression, SPLS can be viewed as a combination of LARS and PLS regression.

The predictive performance of PLSDR and SPLSDR will be compared using generalized ROC curves with the methods of Goeman (Penalized), Sohn et al. (glcoxph), and Park and Hastie (glmpath) using Rosenwald et al. [7] lymphochip microarray with gene expression of patients with diffuse large-B-cell lymphoma. Good predictive performance of the PLSDR algorithm has already been described by Bertrand et al.[8] in 2010. Moreover PLSDR and SPLSDR are very easy to implement under R or other statistical softwares.

Key words: PLS; PLS generalized; PLS Cox; PLSDR; SPLSDR; SPLS; Lasso

Bibliographie

[1] Bastien P., V. Esposito Vinzi, M. Tenenhaus (2005), PLS generalised linear regression, Computational Statistics & Data Analysis 48, 17-46

[2] Bastien P., (2008) Deviance residuals based PLS regression for censored data in hign dimensional setting, Chemometrics and Intelligent Laboratory Systems 91, 78-86
[3] Park M.Y. and Hastie T. (2007) *L*1-regularization path algorithm for generalized linear models *J. R. Statist. Soc.* B 69, *Part* 4, *pp.* 659–677
[4] Sohn J.K., Jung S.H., Park C. (2009) Gradient lasso for Cox proportional hazards model, Bioinformatics, 25(14): 1775-1781.
[5] Goeman J.J. (2010) L1 Penalized Estimation in the Cox proportional Hazards Model, Biometric Journal, 52, 1, 70-84.
[6] Zou H. and Hastie T.(2005) Regularization and Variable Selection via the Elastic Net. *JRSSB* 67(2) 301-320.
[7] Rosenwald A. et al.(2000) the use of molecular profiling to predict survival for diffuse large-B cell lymphoma. The new England Journal of Medicine, 346,1937-1947.
[8] Bertrand F., Maumy-Bertrand M, Beau-Faller M, Meyer N. (2010). plsRcox : modèles de Cox en présence d'un grand nombre de variables explicatives, Congrès Chimiometrie 2010.

Analyse en composantes principales et regression PLS quadratiques

Stéphane Verdun & Véronique Cariou & El Mostafa Qannari

ONIRIS, Unité de Sensométrie et Chimiométrie Nantes, F-44322, France.
INRA, Nantes, F-44316, France.
Université Nantes, Angers, Le Mans, France.

Résumé

Dans le cadre de l'Analyse en Composantes Principales (ACP) et de la régression PLS, plusieurs approches ont été proposées pour tenir compte des relations non linéaires existant entre les variables d'un même tableau ou de deux tableaux de données. Parmi ces approches, nous pouvons citer la régression PLS1 quadratique qui a été proposée par Höskuldsson (1992). Dans cette communication, nous étendrons cette démarche au cadre de la régression PLS2 et de l'Analyse en Composantes Principales. Par la suite, nous présentons des applications notamment dans le cadre de la cartographie des préférences.

Abstract

Within the framework of Principal Components Analysis (PCA) and PLS regression, several approaches have been proposed in order to take account of non linear relationships among variables. Among these approaches, we single out quadratic PLS1 regression which has been proposed by Höskuldsson (1992). We propose herein an extension of quadratic PLS1 regression to the more general framework of PLS2 regression and to Principal Components Analysis. Thereafter, we discuss applications, particularly in the context of preference mapping.

keywords : analyse de données, sensométrie

1 Introduction

L'Analyse en Composantes Principales (ACP) et la régression PLS sont des méthodes très populaires en analyse de données, en sensométrie et en chimiométrie. Elles sont basées sur la détermination de composantes (combinaisons linéaires des variables) qui visent, selon le cas, à synthétiser l'information contenue dans un tableau de données ou à établir un pont entre tableaux de données. Cependant, ces méthodes font l'hypothèse que les relations entre variables sont linéaires. Dans de nombreux cas, cette hypothèse peut être mise en défaut, des relations plus complexes pouvant être mises à jour. C'est pourquoi différentes approches permettant de cerner des relations non linéaires entre variables ont été développées. Dans le cadre de l'ACP et de la régression linéaire, une approche simple

consiste à étendre les tableaux à analyser en rajoutant des transformations des variables (quadratiques, termes croisés,...) (Wold et al., 1989, Berglund et Wold, 1997). Cette approche a néanmoins l'inconvénient de fournir des composantes qui ne sont plus des combinaisons linéaires des variables de départ. Une autre stratégie (Wold et al., 1989, Baffi et al., 1999) consiste à supposer que la relation entre les composantes du tableau X et celles du tableau Y ne sont plus linéaires, mais quadratiques.

Dans le cadre de la régression PLS1, Höskuldsson (1992) a proposé une approche qui consiste à transformer le critère PLS afin de prendre en compte les effets quadratiques et les interactions entre les composantes. Nous proposons d'étendre cette démarche au cadre de la régression PLS2 et de l'analyse en composantes principales. Par la suite, nous présentons une application de cette démarche en cartographie des préférences.

2 PLS quadratique

2.1 PLS1 quadratique

Soit une variable y que nous cherchons à prédire à partir d'un ensemble de variables constituant le tableau X. Dans la suite, toutes les variables sont supposées centrées.

La méthode de PLS quadratique développée par Höskuldsson (1992) consiste à changer le critère de la PLS classique afin de prendre en compte les effets quadratiques et les interactions. De manière plus précise, la n^e composante recherchée est la composante $t_n = Xw_n$ avec $||w_n|| = 1$ qui maximise le critère :

$$C_q = cov^2(y, t_n) + cov^2(y, t_n^2) + cov^2(y, t_n t_1) + \cdots + cov^2(y, t_n t_{n-1})$$

Höskuldsson montre que ce problème peut se résoudre en itérant jusqu'à convergence l'équation :

$$w = \frac{Bw + (w^T Gw)Gw}{w^T Bw + (w^T Gw)^2} \tag{1}$$

avec

$$\begin{array}{rcl} B & = & (\sum y_i x^i)(\sum y_i x^i)^T + \sum_j (y_i t_{ij} x^i)(y_i t_{ij} x^i)^T \\ G & = & X^T diag(y) X \end{array} \tag{2}$$

Une fois la composante t_n déterminée, le rang de la matrice X est réduit par déflation suivant t_n comme cela est fait dans le cadre de la régression PLS1.

2.2 PLS2 quadratique

Dans le cadre de la méthode PLS2, le tableau Y a plusieurs colonnes. Nous recherchons une composante $t = Xw$ associée au tableau X mais aussi une composante $u = Yc$ associée

au tableau Y telles que la somme des covariances carrées entre u et la composante t, de son carré et de ses interactions avec les composantes précédentes soit maximisée :

$$C_q \;=\; cov^2(u,t_n) + cov^2(u,t_n^2) + cov^2(u,t_nt_1) + \cdots + cov^2(u,t_n,t_{n-1})$$

Nous proposons deux algorithmes pour la résolution de ce problème d'optimisation. Le premier algorithme inclut dans la phase itérative de recherche des composantes une phase d'optimisation de la composante u. Le deuxième s'appuie sur un critère alternatif de la PLS2 ne faisant plus intervenir u.

2.3 ACP quadratique

Ayant défini une démarche de régression PLS2 quadratique, il est possible d'en déduire une démarche d'ACP quadratique par application de la régression PLS2 quadratique au cas particulier où $Y = X$. Il est également possible de procéder de manière directe en cherchant à l'étape n une composante t_n de manière à maximiser le critère :

$$\sum_j cov^2(x_j,t_n) + \sum_j cov^2(x_j,t_n^2) + \sum_j \sum_{k<n} cov^2(x_j,t_kt_n)$$

3 La cartographie des préférences

La cartographie des préférences vise à analyser et à expliquer les données de préférences à partir des caractéristiques des produits (sensorielles, physico-chimiques,...) ou à partir des caractéristiques des consommateurs (catégories socio-professionnelles, habitudes de consommations,...). Nous distinguons en particulier la cartographie interne des préférences et la cartographie externe des préférences (Greenhoff et Macfie, 1994).

La cartographie interne des préférences a pour objectif de déterminer les directions de préférences qui sont sous-jacentes à un tableau de données (produits x consommateurs) exprimant les préférences d'un panel de consommateurs pour un ensemble de produits. Dans ce cadre, nous montrons que l'ACP quadratique peut s'avérer un outil pertinent. La cartographie externe des préférences vise à déterminer les directions sous-jacentes à un tableau de données de préférences et, de plus, relier ces directions à des données externes (données sensorielles, par exemple). Dans ce cadre, nous montrons que la démarche PLS quadratique est utile et nous comparons, sur la base de données réelles, ses résultats à ceux qui sont classiquement obtenus à l'aide des démarches usuelles.

Bibliographie

[1] Berglund, A., & Wold, S. (1997) INLR, Implicit non-linear latent variable regression, Journal of Chemometrics, 11, 141-156.

[2] Baffi, G., Martin, E. B., & Morris, A. J. (1999). Non-linear projection to latent structures revisited: the quadratic PLS algorithm. Computers & chemical engineering, 23(3), 393.
[3] Greenhoff, K. and MacFie, H.J.H. (1999) Preference Mapping in practice. In Measurement of food preferences. Editors: H.J.H. MacFie and D.M.H. Thomson. Aspen Publishers, Inc. Gaithersburg, Maryland.
[4] Höskuldsson, A. (1992) Quadratic pls regression, Journal of Chemometrics, 6, 307-334.
[5] Wold, H. (1966) Estimation of principal components and related models by iterative least squares, In Multivariate Analysis, 391-420.
[6] Wold, S., Kettaneh-Wold, N., & Skagerberg, B. (1989) Non-linear PLS modelling, Chemometrics Int. Lab. System, 7, 53-65.

Extrêmes

Nouveaux outils inférentiels pour processus max-stables, *Thomas Opitz, Jean-Noël Barco and Pierre Ribereau*

La modélisation du comportement extrême de phénomènes aléatoires spatiaux est indispensable dans l'évaluation et la gestion des risques environnementaux. L'environnement, la climatologie sont des champs d'application privilégiés (températures, précipitations, neige, marée, vent...). Les champs max-stables spatiaux, en tant que limites de suites normalisées de processus spatiaux, sont des candidats naturels pour modéliser la dépendance spatiale entre événements extrêmes. Nous présentons d'abord la théorie des valeurs extrêmes univariée et multivariée, théorie qui est à la fois fondement et outil pour les champs max-stables. Après avoir défini les processus max-stables et rappelé quelques-unes de leurs propriétés d'intérêt, nous faisons le point sur les principales méthodes inférentielles actuelles. Des outils graphiques originaux et une approche d'inférence innovante pour caractériser la structure de dépendance extrêmale entre paires de sites sont alors proposés.

Estimation semi-paramétrique du paramètre de second ordre en statistique des valeurs extrêmes, *El-Hadji Deme, Laurent Gardes and Stéphane Girard*

Le paramètre d'importance en théorie des valeurs extrêmes est l'indice des valeurs extrêmes. Il contrôle le comportement de la queue de distribution au premier ordre. Plus il est grand, plus la queue est lourde. De nombreux estimateurs de ce paramètre ont été proposés notamment dans le cas particulier où la loi étudiée appartient au domaine d'attraction de Fréchet (cas qui nous intéresse ici). Le plus connu d'entre eux est l'estimateur de Hill (Hill, 1975) qui utilise les k plus grandes observations de l'échantillon. Le biais de ces estimateurs est contrôlé par le paramètre du second ordre. Sa connaissance est donc indispensable lorsqu'il s'agit par exemple de réduire le biais des estimateurs ou encore pour le choix adaptatif du meilleur paramètre k. L'estimation du paramètre du second ordre a fait l'objet de plusieurs études récentes. Citons en particulier (Fraga Alves et al., 2003), (Goegebeur et al., 2010) et enfin (Ciuperca et Mercadier, 2010). Nous proposons un estimateur semi-paramétrique du paramètre du second ordre permettant de regrouper les trois travaux précédents au sein d'un formalisme commun. En particulier, nous montrons qu'il est possible d'établir la normalité asymptotique de ces estimateurs de façon unifiée. Nous tirons également parti de notre formalisme pour proposer de nouveaux estimateurs du paramètre du second ordre.

Estimation d'un paramètre de queue commun aux lois de type Weibull et au domaine d'attraction de Fréchet, *Jonathan El Methni, Laurent Gardes, Stéphane Girard and Armelle Guillou*

Le Théorème de Gnedenko donne les lois limites possibles du maximum d'un échantillon qui sont paramétrèes par l'indice des valeurs extrêmes. Selon son signe elles sont divisées en trois do-

maines d'attraction : Fréchet, Weibull et Gumbel. Dans de nombreuses applications (hydrologie, finance, etc...), les domaines d'attraction de Fréchet et de Gumbel sont très souvent privilégiés. Le domaine d'attraction de Gumbel étant complexe à étudier dans sa globalité, on s'intéresse très souvent à une sous-famille de loi appelée les lois à queue de type Weibull paramètrées par l'indice de queue de Weibull. Afin d'expliquer pourquoi la même méthodologie est utilisée pour estimer l'indice des valeurs extrêmes et l'indice de queue de Weibull, les auteurs proposent une famille de lois regroupant dans un formalisme commun entre autre les lois du domaine d'attraction de Fréchet et les lois à queue de type Weibull. Ces lois dépendent notamment d'un paramètre de forme. Ce paramètre contrôle le comportement de la queue de distribution : plus il est grand plus la queue est lourde et inversement. L'objectif de cette communication est de proposer un estimateur de ce paramètre de forme dont on établira sous certaines hypothèses la loi asymptotique.

Intervalles de confiance pour une fonction implicite des paramètres d'un modèle : application au calcul de l'altitude optimale de présence d'espèces végétales dans une chaîne montagneuse, *Vincent Couallier, Audrey Eyermann, Annabel J. Porté and Magali Urli*

On considère un modèle statistique paramétrique de paramètre $\theta \in \mathbb{R}^p$ et on se place dans le cadre d'une estimation par maximum de vraisemblance. Dans cette note, on propose une méthode de calcul d'intervalle de confiance pour une quantité $\phi \in \mathbb{R}$ qui est définie implicitement à partir de θ de la façon suivante : soit g une fonction continue à dérivée continue qui permet de définir implicitement $\phi \in \mathbb{R}^q$ par rapport à θ : $g(\phi, \theta) = 0$, en supposant que ϕ est défini de façon unique.

L'application concerne la détermination de l'altitude optimale de présence d'espèces végétales le long d'un gradient d'altitude dans une chaine montagneuse.

Nouveaux outils inférentiels pour processus max-stables

Thomas Opitz & Jean-Noël Bacro & Pierre Ribereau

Institut de Mathématique et de Modélisation (UMR 5149)
Université Montpellier 2, Case Courrier 051, Place Eugène Bataillon, F-34095 Montpellier

Résumé. La modélisation du comportement extrême de phénomènes aléatoires spatiaux est indispensable dans l'évaluation et la gestion des risques environnementaux. L'environnement, la climatologie sont des champs d'application privilégiés (températures, précipitations, neige, marée, vent ...). Les champs max-stables spatiaux, en tant que limites de suites normalisées de processus spatiaux, sont des candidats naturels pour modéliser la dépendance spatiale entre événements extrêmes. Nous présentons d'abord la théorie des valeurs extrêmes univariée et multivariée, théorie qui est à la fois fondement et outil pour les champs max-stables. Après avoir défini les processus max-stables et rappelé quelques-unes de leurs propriétés d'intérêt, nous faisons le point sur les principales méthodes inférentielles actuelles. Des outils graphiques originaux et une approche d'inférence innovante pour caractériser la structure de dépendance extrêmale entre paires de sites sont alors proposés.

Mots-clés : théorie des valeurs extrêmes ; processus max-stables ; modélisation spatiale

Abstract. Modelling the extreme behaviour of random spatial phenomena is indispensable for the management of environmental risks. Primary fields of application are environment and climatology (temperatures, precipitation, snowfall, storm tides/surges, wind ...). Max-stable fields, which arise as limits of normalised sequences of spatial processes, are obvious candidats for the modelisation of spatial dependence between extreme events. First we present univariate and multivariate extreme value theory which is at the same time foundation and tool for max-stable processes. Then we define max-stable processes and recapitulate some of their properties of interest, together with the current principal inferential methods. Finally we propose novel graphical tools and a new inference approach to characterize the extremal dependence structure between site pairs.
Keywords : Extreme Value Theory ; max-stable processes ; spatial modelisation

1 Introduction

Les applications statistiques de la théorie des valeurs extrêmes supposent souvent une prédiction au-delà des plus grandes observations, c'est-à-dire une extrapolation. Les résultats fondamentaux de la théorie concernent donc, de façon naturelle, le cadre asymptotique. Une certaine régularité dans la queue des distributions est nécessaire pour obtenir asymptotiquement une loi limite non-dégénérée. En pratique, on fait l'hypothèse que cette régularité est présente dans les observations, ou en d'autres termes que l'on est assez loin dans la queue pour avoir une bonne approximation du comportement asympotique.
Pour plus de détails sur les rappels qui suivent, nous renvoyons à la monographie de S. Resnick [4] qui couvre surtout les aspects théoriques.

2 Extrêmes univariés

2.1 Maxima

Soient $X, X_1, ..., X_n$ variables aléatoires iid, $X \sim F$. On suppose l'existence de suites normalisantes $a_n > 0$, b_n telles que

$$\frac{\max\limits_{i=1,...,n} X_i - b_n}{a_n} \xrightarrow{d} Z \quad (n \to \infty) \,, \tag{1}$$

où Z est une variable aléatoire non-dégénérée de loi G. On dit alors que F appartient au domaine d'attraction de G, noté $F \in \mathcal{D}(G)$. La loi G de Z est nécessairement max-stable, c'est-à-dire qu'il existe des fonctions $\alpha(t), \beta(t)$ telles que $G^t(\alpha(t)x + \beta(t)) = G(x)$ pour $t > 0$. Les lois max-stables univariées peuvent être paramétrisées de la façon suivante :

$$G(x) = G_{\xi,\mu,\sigma}(x) = \exp\left(-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right) \,.$$

Cette écriture, dites *loi généralisée des extrêmes*, comporte trois paramètres : un paramètre de forme ξ, caractéristique de la queue de distribution, un paramètres de localisation μ et un d'échelle $\sigma > 0$. La loi Fréchet unité $G(x) = \exp\left(-\frac{1}{x}\right)$ est obtenue en posant $\xi = \mu = \sigma = 1$.
Dans ce qui suit, nous utilisons la notation $\widetilde{X} = \frac{X - b_n}{a_n}$.

2.2 Dépassement de seuils

Soit $F \in \mathcal{D}(G)$. Lorsque $n \to \infty$, les dépassement de seuils $Y = \widetilde{X} - u \mid \widetilde{X} > u$ convergent en loi vers une loi de Paréto généralisée (GPD), de fonction de répartition

$$F_{\text{GPD}}(y) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-\frac{1}{\xi}} \,,$$

où $\sigma_u = \sigma + \xi u$.

2.3 Processus de Poisson

On peut (facilement) montrer que les deux représentations précedentes du comportement asymptotique sont équivalentes à

$$nP(\widetilde{X} \in \cdot) \Rightarrow \mu(\cdot) \quad (n \to \infty), \tag{2}$$

avec $\mu(]z, \infty[) = (1 + \xi\frac{z-\mu}{\sigma_u})_+^{-\frac{1}{\xi}}$.
En définissant $N_n = \left\{\widetilde{X}_i \mid i = 1, 2, ..., n\right\}$, (2) nous amène à la convergence faible $N_n \Rightarrow N$ $(n \to \infty)$ avec un processus de Poisson N de mesure d'intensité μ.

3 Extrêmes multivariés

Soient $\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}, ...$ des vecteurs aléatoires iid t.q. $\mathbf{X} = (X^{(1)}, ..., X^{(d)}) \sim F(F_1, ..., F_d)$, avec les fonctions de répartitions marginales F_i de $\mathbf{X}$.
La théorie des extrêmes univariés s'applique aux marginales univariées et les notions présentées précédemment se généralisent au cadre multivarié.

⚠ Les notations suivantes sont vectorielles, et les opérations entre vecteurs se font par composante des vecteurs.

Soient $\mathbf{a_n}, \mathbf{b_n}$ vecteurs normalisants tels que

$$\frac{\max\limits_{i=1,...,n} \mathbf{X_i} - \mathbf{b_n}}{\mathbf{a_n}} \xrightarrow{d} \mathbf{Z} \sim G(G_1, ..., G_d) \ .$$

De nouveau, G est max-stable : $G^t(\boldsymbol{\alpha}(\mathbf{t})\mathbf{x} + \boldsymbol{\beta}(\mathbf{t})) = G(\mathbf{x})$ pour tout $t > 0$ et des vecteurs $\boldsymbol{\alpha}(\mathbf{t})$, $\boldsymbol{\beta}(\mathbf{t})$. De plus, les marginales G_i sont des lois extrêmes univariées.
G permet la représentation

$$G(x) = \exp\left(-\mu([-\boldsymbol{\infty}, \mathbf{x}]^c)\right) \ .$$

avec la *mesure exponent* μ, une mesure de Radon soumise à certaines contraintes.
Soit le processus ponctuel

$$N_n = \left\{ \widetilde{\mathbf{X_i}} \mid i = 1, ..., n \right\} \ ,$$

alors

$$N_n \Rightarrow N \ ,$$

où N est un processus de Poisson de mesure d'intensité μ.

Sans perte de généralité on peut toujours se ramener à des marginales Fréchet unité (notées G_i^*). La mesure exponent $\mu = \mu^*$, définie sur $[0, \infty]^d$, vérifie alors une propriété d'homogénéité d'ordre -1 : $\mu^*(tB) = t^{-1}\mu^*(B)$ pour tout B ensemble de Borel. Cette propriété permet une factorisation de μ^* à l'aide de la *mesure spectrale*.
Pour ce faire, choisissons deux normes $\|\cdot\|_a$, $\|\cdot\|_r$ afin de définir la transformation vers des *coordonnées angulaires*

$$T : \mathbb{R}_+^d \to \mathbb{R}_+ \times S_+, \quad T(\mathbf{x}) = \left(\|\mathbf{x}\|_r, \frac{\mathbf{x}}{\|\mathbf{x}\|_a} \right),$$

La mesure spectrale ρ est une mesure de Radon finie sur S_+, et on obtient

$$\mu^*(d(r, a)) = r^{-2} dr \times \rho(da) \ .$$

Il n'y a pas de paramétrisation finie pour la mesure spectrale. Elle caractérise la structure de dépendance extrêmale d'un vecteur aléatoire dans le domaine d'attraction d'une loi extrême.

4 Champs max-stables

La max-stabilité d'un champ $\{Z(x) \mid x \in \mathbb{R}^d\}$ se définit par la max-stabilité de ses lois fini-dimensionnelles. Comme avant, les champs max-stables se définissent comme processus limites pour une suite de champs aléatoires iid, sous normalisation linéaire par une suite de champs déterministe $a_n(x)$, $b_n(x)$.

4.1 Construction

Nous mentionnons une méthode pour construire des champs max-stables qui est à la fois une caractérisation (voir [1] et [5] pour plus de détails).
Soient (V_i, U_i) les points d'un processus de Poisson de mesure d'intensité $v^{-2}dv \times du$ sur $]0, \infty[\times\mathbb{R}^d$, et soit $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ une fonction non-négative telle que $\int f(x, y)dy = 1$ pour tout x. Alors

$$Z(x) = \max_{i=1,2,...} (V_i f(x, U_i))$$

définit un champs max-stable sur $\mathbb{R}^d$ avec des marginales Fréchet unité.
En prenant f comme la densité d'une loi normale multivariée, on obtient un processus max-stable dit *processus gaussien de valeurs extrêmes* ([5]).

4.2 Méthodes et outils inférentiels

L'inférence est basée sur les lois fini-dimensionnelles, en particulier sur les lois bivariées pour deux sites spatiaux à distance Δx. Une statistique scalaire résumant la structure de dépendance est le coefficient extrêmal $\theta(\Delta x) = -\log G(1, 1; \Delta x)$, qui est situé entre 1 (dépendance complète) et 2 (indépendance).
Les outils de géostatistique comme par exemple le variogramme incitent à utiliser des équivalents adaptés au cadre max-stable. Le madogramme, le F-madogramme et le λ-madogramme tiennent compte du fait que les seconds moments des marginales n'existent pas en général (voir par exemple [2]).
Les méthodes d'inférence usuelles sont essentiellement fondées sur l'ajustement de versions paramétriques du coefficient extrêmal ou sur des approches par maximum de vraisemblance composite. Le plus souvent, seules les expressions analytiques des lois bivariées sont disponibles, ce qui rend toute approche par maximum de vraisemblance classique impossible. Récemment [3] ont proposé une approche par vraisemblance composite sur les paires de maxima par période. L'approche que nous proposons cherche quant à elle à exploiter des propriétés de la mesure spectrale et les processus limites de Poisson des extrêmes multivariés. Différents résultats théoriques sont présentés, et des approches orientées "dépassements de seuil" ou "plus grandes statistiques d'ordre" sont proposées. De nouveaux outils graphiques et des méthodes inférentielles paramétriques innovantes, respectant la structure spatiale ou spatio-temporelle des observations extrêmes sont explicités.

Références

[1] L. De Haan. A spectral representation for max-stable processes. *The Annals of Probability*, 12(4) :1194–1204, 1984.

[2] P. Naveau, A. Guillou, D. Cooley, and J. Diebolt. Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1) :1, 2009.

[3] S.A. Padoan, M. Ribatet, and S.A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489) :263–277, 2010.

[4] S.I. Resnick. *Extreme values, regular variation and point processes*. Springer Verlag, 2007.

[5] R.L. Smith. Max-stable processes and spatial extremes. *Unpublished manuscript*, 205, 1990.

Estimation semi-paramétrique du paramètre de second ordre en statistique des valeurs extrêmes

El-hadji Deme[1,2], Laurent Gardes[1] & Stéphane Girard[1]

[1] *Equipe Mistis, INRIA Rhône-Alpes & Laboratoire Jean Kuntzmann*
655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier Cedex, France.
Stephane.Girard@inrialpes.fr

[2] *Université Gaston Berger de Saint-Louis*
BP 234, Saint-Louis, Sénégal.

Résumé Le paramètre d'importance en théorie des valeurs extrêmes est l'indice des valeurs extrêmes noté γ. Il contrôle le comportement de la queue de distribution au premier ordre. Plus γ est grand, plus la queue est lourde. De nombreux estimateurs de ce paramètre ont été proposés notamment dans le cas particulier où la loi étudiée appartient au domaine d'attraction de Fréchet (cas qui nous intéresse ici). Le plus connu d'entre eux est l'estimateur de Hill (Hill, 1975) qui utilise les k plus grandes observations de l'échantillon. Le biais de ces estimateurs est contrôlé par le paramètre du second ordre ρ. La connaissance de ρ est donc indispensable lorsqu'il s'agit par exemple de réduire le biais des estimateurs ou encore pour le choix adaptatif du meilleur paramètre k. L'estimation de ce paramètre a fait l'objet de plusieurs études récentes. Citons en particulier (Fraga Alves *et al.*, 2003), (Goegebeur *et al.*, 2010) et enfin (Ciuperca et Mercadier, 2010). Nous proposons un estimateur semi-paramétrique de ρ permettant de regrouper les trois travaux précédents au sein d'un formalisme commun. En particulier, nous montrons qu'il est possible d'établir la normalité asymptotique de ces estimateurs de façon unifiée. Nous tirons également parti de notre formalisme pour proposer de nouveaux estimateurs de ρ.

Mots clés : Théorie des valeurs extrêmes, paramètre du second ordre, Estimateur semi-paramétrique.

Abstract An important parameter in extreme value theory is the extreme value index γ. It controls the first order behavior of the distribution tail. In the literature, numerous estimators of this parameter have been proposed especially in the case of heavy tail distribution (which is the situation considered here). The most known estimator was proposed by (Hill, 1975). It depends on the k largest observations of the underlying sample. The bias of the tail index estimator is controlled by the second order parameter ρ. In order to reduce the bias of γ's estimators or to select the best number k of observations to use, the knowledge of ρ is essential. Some estimators of ρ can be found in the literature, see for example (Fraga Alves *et al.*, 2003), (Goegebeur *et al.*, 2010) and (Ciuperca et Mercadier, 2010). We propose a semiparametric estimator of ρ that encompasses the three previously mentioned estimators. The asymptotic normality of these estimators is then proved in an

unified way. New estimators of ρ are also introduced.

Keywords : Extreme Value Theory, second order parameter, semiparametric estimator.

1 Statistique des valeurs extrêmes

Nous nous intéressons à la situation où la fonction de répartition F d'un échantillon $(X_1, \ldots, X_n)$ appartient au domaine d'attraction de Fréchet, *i.e.*

(B1) Il existe $\gamma > 0$ tel que :

$$\lim_{x\to\infty} \frac{1 - F(\lambda x)}{1 - F(x)} = \lambda^{-1/\gamma}, \ \forall \lambda > 0.$$

Cette condition signifie que $1 - F$ est à variations régulières à l'infini d'indice γ. Il est équivalent de supposer que

$$U(x) := F^{\leftarrow}(1 - 1/x) = x^{\gamma}\ell(x),$$

où $F^{\leftarrow}$ est l'inverse généralise de F et ℓ est une fonction à variations lentes à l'infini *i.e.*

$$\lim_{x\to\infty} \frac{\ell(\lambda x)}{\ell(x)} = 1.$$

En d'autres termes, la fonction de répartition F est dite à queue lourde. Le paramètre γ est appelé indice des valeurs extrêmes. Il gouverne le comportement au premier ordre de la queue de la loi. Plus γ est grand, plus la queue est lourde. L'estimation de γ a fait l'objet de nombreuses études. Dekkers *et al.* (1989) ont proposé un estimateur basé sur les moments empiriques :

$$\mathcal{M}_n^{(\alpha)}(k) = \frac{1}{k}\sum_{j=1}^{k}(\log X_{n-j+1,n} - \log X_{n-k,n})^{\alpha}, \ \alpha > 0,$$

où $X_{1,n} \leq \ldots \leq X_{n,n}$ sont les statistiques d'ordre triées par ordre croissant et où $1 < k < n$. Beirlant *et al.* (1999) ont montré que les variables aléatoires

$$\mathcal{S}_n^{(H)}(k) = \frac{1}{k}\sum_{j=1}^{k} H\left(\frac{j}{k+1}\right) j(\log X_{n-j+1,n} - \log X_{n-j,n}),$$

peuvent également être utilisées pour estimer γ. Ici, H est un noyau vérifiant certaines propriétés. Plus récemment, Ciuperca *et al.* (2010) ont étudié l'estimation de l'indice des valeurs extrêmes à partir des moments empiriques pondérés :

$$\mathcal{WM}_n^{(g,\alpha)}(k) = \frac{1}{k}\sum_{j=1}^{k} g\left(\frac{j}{k+1}\right)(\log X_{n-j+1,n} - \log X_{n-k,n})^{\alpha}, \ \alpha > 0,$$

où g est une fonction poids positive. Afin d'établir les propriétés asymptotiques de ces estimateurs de l'indice des valeurs extrêmes, une condition du second ordre est requise :

(B2) Il existe une fonction $A(x) \to 0$ de signe constant à l'infini et un paramètre du second ordre $\rho \leq 0$ tels que, pour tout $\lambda > 0$,

$$\lim_{x\to\infty} \frac{\log U(\lambda x) - \log U(x) - \gamma \log \lambda}{A(x)} = \int_1^\lambda u^{\rho-1} du.$$

Remarquons que la condition **(B2)** implique que $|A|$ est une fonction à variations régulières d'indice ρ, voir Geluk *et al.* (1987). Par conséquent, plus le paramètre du second ordre ρ est petit, plus la vitesse de convergence dans **(B1)** est grande. La connaissance de ρ est alors d'un grand intérêt dans la pratique. Par exemple, le paramètre du second ordre est de première importance dans le choix adaptatif du meilleur paramètre k lors de l'estimation de l'indice des valeurs extrêmes.

2 Estimation du paramètre du second ordre

L'estimation du paramètre du second ordre ρ a fait l'objet de plusieurs études récentes. Citons en particulier (Fraga Alves *et al.*, 2003) qui utilisent les statistiques $\mathcal{M}_n^{(\alpha)}(k)$, (Goegebeur *et al.*, 2010) qui se basent sur $\mathcal{S}_n^{(H)}(k)$ et enfin (Ciuperca et Mercadier, 2010) qui considèrent les statistiques $\mathcal{WM}_n^{(g,\alpha)}(k)$.

Nous proposons un estimateur semi-paramétrique de ρ permettant de regrouper les trois travaux précédents au sein d'un formalisme commun. En particulier, nous montrons qu'il est possible d'établir la normalité asymptotique de ces estimateurs de façon unifiée en se basant sur le condition classique du troisième ordre :

(B3) Il existe des fonctions $A(x) \to 0$ et $B(x) \to 0$ toutes deux de signe constant à l'infini, un paramètre du second ordre $\rho \leq 0$ et un paramètre du troisième ordre $\beta \leq 0$ tels que, pour tout $\lambda > 0$,

$$\lim_{x\to\infty} \frac{\dfrac{\log U(\lambda x) - \log U(x) - \gamma \log \lambda}{A(x)} - K_\rho(\lambda)}{B(x)} = \int_1^\lambda s^{\rho-1} \int_1^s u^{\beta-1} du ds$$

et où il peut être montré que $|B|$ est à variations régulières d'indice β.

Finalement, nous tirons également parti de notre formalisme pour proposer de nouveaux estimateurs de ρ.

Références

[1] J. Beirlant, G. Dierckx, Y. Goegebeur and G. Matthys. (1999) Tail index estimation and an exponential regression model. *Extremes*, **2**, 177–200.

[2] G. Ciuperca and C. Mercadier. (2010) Semi-parametric estimation for heavy tailed distributions. *Extremes*, **13**, 55–87.

[3] A.L.M. Dekkers, J.H.J. Einmahl, and L. de Haan. (1989) A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, **17**, 1833–1855.

[4] M.I. Fraga Alves, M.I. Gomes, and L. de Haan. (2003) A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica*, **60**(2), 193–213, 2003.

[5] J. Geluk and L. de Haan. (1987) *Regular Variation, Extensions and Tauberian Theorems*, CWI Tract 40, Center for Mathematics and Computer Science, Amsterdam, Netherlands, 1987.

[6] Y. Goegebeur, J. Beirlant, and T. de Wet. (2010) Kernel estimators for the second order parameter in extreme value statistics. *Journal of Statistical Planning and Inference*, **140**, 2632–2652.

[7] B.M. Hill. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

Estimation d'un paramètre de queue commun aux lois de type Weibull et au domaine d'attraction de Fréchet

Jonathan El Methni[(1)], Laurent Gardes[(1)], Stéphane Girard[(1)] & Armelle Guillou[(2)]

[(1)] *Equipe Mistis, INRIA Rhône-Alpes & Laboratoire Jean Kuntzmann*
655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.

[(2)] *Université de Strasbourg & CNRS, IRMA, UMR 7501,*
7, rue René Descartes, 67084 Strasbourg cedex, France.

jonathan.el-methni@inrialpes.fr

Résumé

Le Théorème de Gnedenko [8] donne les lois limites possibles du maximum d'un échantillon qui sont paramétrées par l'indice des valeurs extrêmes γ. Selon le signe de γ elles sont divisées en trois domaines d'attraction : Fréchet ($\gamma > 0$), Weibull ($\gamma < 0$) et Gumbel ($\gamma = 0$). Dans de nombreuses applications (hydrologie, finance, etc ...), les domaines d'attraction de Fréchet et de Gumbel sont très souvent privilégiés. Le domaine d'attraction de Gumbel étant complexe à étudier dans sa globalité, on s'intéresse très souvent à une sous-famille de loi appelée les lois à queue de type Weibull paramétrées par l'indice de queue de Weibull $\tilde{\theta}$. Afin d'expliquer pourquoi la même méthodologie est utilisée pour estimer ces paramètres γ et $\tilde{\theta}$, les auteurs proposent dans [6], une famille de lois regroupant dans un formalisme commun entre autre les lois du domaine d'attraction de Fréchet et les lois à queue de type Weibull. Ces lois dépendent de 2 paramètres $\tau \in [0, 1]$ et $\theta > 0$. Le cas $\tau = 0$ correspond à une loi à queue de type Weibull et le cas $\tau = 1$ correspond à une loi appartenant au domaine d'attraction de Fréchet. L'objectif de cette communication est de proposer un estimateur de τ indépendant de θ. Ce paramètre contrôle le comportement de la queue de distribution : plus il est grand plus la queue est lourde et inversement. On établira sous certaines hypothèses la loi asymptotique de l'estimateur de τ.

Mots clés : Loi de type Weibull, normalité asymptotique, statistique des valeurs extrêmes.

Abstract

The Gnedenko Theorem [8] is a general result in extreme value theory regarding the asymptotic distribution of extreme order statistics. The maximum of a sample of *iid* random variables after proper renormalization converges in distribution to one of 3 possible maximum domain of attraction : Fréchet ($\gamma > 0$), Weibull ($\gamma < 0$) and Gumbel ($\gamma = 0$). In a lot of applications (hydrology, finance, etc ...), the Fréchet maximum domain of attraction and the Gumbel maximum domain of attraction are used. The Gumbel maximum domain of attraction encompasses a large variety of distributions, which is why we are interested in a subfamily of distributions called Weibull tail-distributions which depend

on the Weibull tail-coefficient $\tilde{\theta}$. In order to explain why the same methodology is used to estimate the parameters γ and $\tilde{\theta}$, the authors propose in [6], a family of distributions which encompasses the whole Fréchet maximum domain of attraction as well as Weibull tail-distributions. These distributions depend on 2 parameters $\tau \in [0,1]$ and $\theta > 0$. The first parameter τ allows us to represent a large panel of distribution tails ranging from Weibull-type tails ($\tau = 0$) to distributions belonging to the maximum domain of attraction of Fréchet ($\tau = 1$). The main goal of this communication is to propose an estimator of τ that is independent of θ. This parameter controls the behavior of the tail-distribution : the larger the value of τ, the heavier is the tail. Under some assumptions we establish the asymptotic distribution of the estimator.

Keywords : Weibull tail-distributions, asymptotic normality, extreme-value statistics.

1 Introduction

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F. On note $X_{1,n} \leq \ldots \leq X_{n,n}$ l'échantillon ordonné associé. Le Théorème de Gnedenko donne les lois limites possibles du maximum d'un échantillon :

Sous des hypothèses générales sur F il existe deux suites de normalisation $a_n > 0$ et b_n et un réel γ tels que :

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = H_\gamma(x),$$

avec $H_\gamma(x) = \exp(-(1+\gamma x)_+^{-\frac{1}{\gamma}})$ si $\gamma \neq 0$ et $H_0(x) = \exp(-e^{-x})$, où $y_+ = \max(y, 0)$.

On dit que H_γ est la loi des valeurs extrêmes et γ l'indice des valeurs extrêmes.

Selon le signe de γ on distingue 3 domaines d'attraction (D.A.) :

- Si $\gamma > 0$, F appartient au D.A. de Fréchet. Ce sont les lois à queue lourde (Pareto, Student).
- Si $\gamma < 0$, F appartient au D.A. de Weibull. Ce sont les lois dont le point terminal est fini (Beta, uniforme).
- Si $\gamma = 0$, F appartient au D.A. de Gumbel. Ce sont les lois dont la fonction de survie décroit vers zéro à une vitesse exponentielle (exponentielle, normale, gamma, Weibull).

Dans de nombreuses applications (hydrologie, finance, etc ...), les D.A. de Fréchet et de

Gumbel sont très souvent privilégiés. Concernant le D.A. de Fréchet, il existe de nombreux estimateurs de γ, le plus connu étant l'estimateur de Hill [9] :

$$H_n(k_n) = \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})),$$

où $1 < k_n < n$ est une suite d'entiers. Le D.A. de Gumbel étant complexe à étudier dans sa globalité, on s'intéresse à une sous-famille de loi appelée les lois à queue de type Weibull dont la fonction de survie $\bar{F} = 1 - F$ vérifie :

$$\exists \tilde{\theta} > 0 \quad \text{tel que} \quad \forall \lambda > 0 \quad \lim_{t \to +\infty} \frac{\log \bar{F}(\lambda t)}{\log \bar{F}(\lambda)} = \lambda^{1/\tilde{\theta}}.$$

Autrement dit $\log \bar{F}$ est à variations régulières d'indice $1/\tilde{\theta}$. Le paramètre $\tilde{\theta}$ est le coefficient de queue de Weibull. L'estimation de $\tilde{\theta}$ fait aussi l'objet d'une grande littérature citons par exemple : [1, 2, 3, 4, 5, 6, 7]. Le point commun entre les estimateurs de $\gamma > 0$ et $\tilde{\theta}$ est qu'ils sont basés sur la différence entre les log (log-spacings) des k_n statistiques d'ordres supérieures. Dans [6], les auteurs proposent une famille de lois regroupant dans un formalisme commun entre autre les lois du D.A. de Fréchet et les lois à queue de type Weibull. Plus précisement ils considèrent les fonctions de survie $\bar{F}$ définies par :

$$(A_1(\tau, \theta)) \quad \bar{F}(x) = \exp(-K_\tau^{\leftarrow}(\log(H(x)))) \quad \text{pour} \quad x \geq x_* \quad \text{avec} \quad x_* > 0$$

et :

$$K_\tau(x) = \int_1^x u^{\tau - 1} du \quad \text{avec} \quad \tau \in [0, 1]$$

La fonction H est croissante et telle que son inverse généralisée $H^{\leftarrow}(t) = \inf(x, H(x) \geq t) = t^\theta \ell(t)$, où $\theta > 0$ et ℓ est une fonction à variations lentes c'est à dire telle que, $\forall t > 0$,

$$\lim_{x \to +\infty} \frac{\ell(tx)}{\ell(x)} = 1 \tag{1}$$

La fonction $H^{\leftarrow}$ est dite à variations régulières d'indice θ. Le cas $\tau = 0$ correspond à une loi à queue de type Weibull et le cas $\tau = 1$ correspond à une loi appartenant au domaine d'attraction de Fréchet. Dans [6], les auteurs proposent d'estimer θ par :

$$\hat{\theta_n}(k_n) = \frac{H_n(k_n)}{\mu_\tau\left(\log(\frac{n}{k_n})\right)} \tag{2}$$

où,

$$\forall t > 0, \quad \mu_\tau(t) = \int_0^\infty (K_\tau(v + t) - K_\tau(t)) e^{-v} dv.$$

Ils établissent la normalité asymptotique de $\hat{\theta}_n(k_n)$ en introduisant la condition du second ordre :

$(A_2(\rho))$ Il existe $\rho < 0$ et $b(x) \to 0$ tels que uniformément localement sur $\lambda \geq \lambda_0 \geq 0$ on ait :

$$\log\left(\frac{\ell(\lambda x)}{\ell(x)}\right) \sim b(x) K_\rho(\lambda) \quad \text{pour} \quad x \to +\infty$$

où $|b|$ est une fonction à variations régulières d'indice ρ asymptotiquement décroissante.

Cette condition permet de contrôler la vitesse de convergence dans (1). L'estimateur de θ proposé dépend de τ. L'objectif de cette communication est de proposer un estimateur de τ indépendant de θ. Le paramètre τ que l'on cherche à estimer permet de contrôler le comportement de la queue de distribution : plus il est grand plus la queue est lourde et inversement.

2 Définition de l'estimateur et sa loi asymptotique

On propose d'estimer τ par :

$$\hat{\tau}_n = \begin{cases} {\psi_n}^{-1}\left(\log\left(\frac{H_n(k_n)}{H_n(k_n')}\right)\right) & \text{si} \quad \frac{H_n(k_n)}{H_n(k_n')} < \frac{k_n'}{k_n} \\ U & \text{si} \quad \frac{H_n(k_n)}{H_n(k_n')} \geq \frac{k_n'}{k_n}. \end{cases}$$

où,

$$\psi_n(x) = \log\left(\frac{\mu_x(\log(n/k_n))}{\mu_x(\log(n/k_n'))}\right).$$

et U est la réalisation d'une loi uniforme sur $[0,1]$. L'estimateur $\hat{\tau}_n$ existe car on peut montrer que la fonction ψ_n est bijective de $\mathbb{R}$ dans $]-\infty, \log(k_n'/k_n)[$.

Théorème 1 *On se place sous les hypothèses $(A_1(\tau,\theta))$ et $(A_2(\rho))$. Soient deux suites d'entiers (k_n) et (k_n') telles que :*

$$k_n \to +\infty, \quad k_n'/n \to 0, \quad k_n/k_n' \to 0, \quad \sqrt{k_n'} b(\exp K_\tau(\log n/k_n')) \to 0.$$

$$\sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right) / \log_2(n/k_n) \to +\infty$$

$$\log(n/k_n')\left(\log_2(n/k_n) - \log_2(n/k_n')\right) \to +\infty.$$

On a :

$$\sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right)(\hat{\tau}_n - \tau) \xrightarrow{d} \mathcal{N}(0,1).$$

Dans un travail futur, on se propose d'étudier le comportement asymptotique de l'estimateur de θ (2) en remplacant τ par $\hat{\tau}_n$.

Références

[1] M. Broniatowski. (1993). On the estimation of the Weibull tail coefficient. *Journal of Statistical Planning and Inference*, **35**, 349–366.

[2] J. Diebolt, L. Gardes, S. Girard and A. Guillou. (2009). Bias-reduced estimators of the Weibull tail-coefficient. *Test*, **17**, 311–331.

[3] G. Dierckx, J. Beirland, D. De Waal and A. Guillou. (2008). A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*, **139**, 1905–1920.

[4] L. Gardes and S. Girard. (2006). Comparison of Weibull tail-coefficient estimators. *REVSTAT-Statistical Journal*, **4**, 163–188.

[5] L. Gardes and S. Girard. (2008). Estimation of the Weibull tail-coefficient with linear combinaison of upper order statistics. *Journal of Statistical Planning and Inference*, **138**, 1416–1427.

[6] L. Gardes, S. Girard and A. Guillou. (2011). Weibull tail-distributions revisited : a new look at some tail estimators. *Journal of Statistical Planning and Inference*, **141**, 429–444

[7] S. Girard. (2004). A Hill type estimate of the Weibull tail-coefficient. *Communication in Statistics- Theory and Methods*, **33**(2), 205–234.

[8] B.V. Gnedenko. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, **44**, 423–453.

[9] B.M. Hill. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, **3**, 1163–1174.

Intervalles de confiance pour une fonction implicite des paramètres d'un modèle : application au calcul de l'altitude optimale de présence d'espèces végétales dans une chaine montagneuse.

Vincent Couallier[1] & Audrey Eyermann[1] & Annabel J. Porté[2,3] & Morgane Urli[2,3]

(1) Institut Mathématique de Bordeaux UMR CNRS 5251, Université Victor Segalen
(2) INRA, UMR 1202 BIOGECO, F-33610 Cestas, France
(3) Université de Bordeaux, UMR 1202 BIOGECO, F-33400 Talence, France

Résumé : En couplant le théorème des fonctions implicites et la "méthode delta", on propose une méthode d'estimation de la variance d'un estimateur d'un paramètre défini implicitement comme le zéro d'une fonction $g(.,\theta)$ où θ est estimé par maximum de vraisemblance. L'application concerne la détermination de l'altitude optimale de présence d'espèces végétales le long d'un gradient d'altitude dans une chaine montagneuse.

Summary : By using a theorem of implicit functions and the Delta Méthod, we propose a method to estimate the variance of an estimator of an implicitly defined parameter such that the solution of $g(.,\theta) = 0$ where θ is estimated by the maximum likelihood method. We apply this to determine the altitude of the optimum of species presence on an altitudinal gradient in mountains.

Mots clés : fonctions implicites, intervalles de confiance, maximum de vraisemblance, delta méthode, optimum d'altitude.

1 Méthodes de calculs d'intervalles de confiance pour des fonctions implicites de paramètres d'un modèle statistique

On considère un modèle statistique paramétrique de paramètre $\theta \in \mathbb{R}^p$ et on se place dans le cadre d'une estimation par maximum de vraisemblance. Dans cette note, on propose une méthode de calcul d'intervalle de confiance pour une quantité $\phi \in \mathbb{R}$ qui est définie implicitement à partir de θ de la façon suivante : soit g une fonction continue à dérivée continue qui permet de définir implicitement $\phi \in \mathbb{R}^q$ par rapport à θ : $g(\phi, \theta) = 0$, en supposant que ϕ est défini de façon unique.

L'estimateur du maximum de vraisemblance du paramètre θ possède de nombreuses propriétés d'optimalité sous des hypothèses standarts. A partir de la normalité asymptotique de $\hat{\theta}_n$, l'objectif est d'obtenir la normalité (asymptotique) de $\hat{\phi}_n$ qui est solution

de $g(\hat{\phi}_n, \hat{\theta}_n) = 0$. De ce résultat, on peut déduire une construction d'intervalle de confiance. Cette méthode est comparée à un calcul d'intervalle de confiance par bootstrap non paramétrique.

1.1 Théorème des fonctions implicites et Delta-méthode

La "méthode delta" (voir Billingsley 1986, Serfling, 1980) permet de déduire d'un résultat de normalité (asymptotique le plus souvent) d'un estimateur $\hat{\theta}_n \in \mathbb{R}^p$, une approximation en loi d'une statistique de type $\hat{\phi}_n = f(\hat{\theta}_n)$, où $\hat{\phi}_n \in \mathbb{R}^q$, f est une fonction continuement différentiable de $\mathbb{R}^p$ dans $\mathbb{R}^q$. Si on connaît le gradient de f, évalué au point $\hat{\theta}_n$, et si on a un résultat du type

$$\Sigma_n^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, Id_p)$$

où Σ_n est la matrice de variance asymptotique de $\hat{\theta}_n$, alors

$$(\nabla f' Var \nabla f)^{-1/2}_{|\theta=\hat{\theta}_n}(\hat{\phi}_n - \phi) \rightarrow_d \mathcal{N}(0, Id_q).$$

Couramment employée pour calculer la variance d'une transformation non linéaire des paramètres à partir de la normalité asymptotique d'un estimateur du maximum de vraisemblance, cette relation nécessite évidemment de connaître ou d'estimer la variance de $\hat{\theta}_n$ mais aussi d'expliciter le gradient de f. Or il existe des modèles où la quantité d'intérêt ϕ est définie implicitement à partir du paramètre θ, par exemple comme la racine d'une équation. L'objectif est donc de proposer, par une utilisation du théorème des fonctions implicites, une représentation du résultat ci-dessus dans le cas où ϕ est défini par $g(\phi, \theta) = 0$ où g est une fonction de $\mathbb{R}^{p+q}$ dans $\mathbb{R}^q$.

Le théorème des fonctions implicites (Taylor et Mann, 1983), affirme, sous certaines conditions, l'existence de fonctions continuement dérivables à valeurs réelles $f_1, \ldots, f_q$ dont on connait les dérivées partielles et qui sont telles que $g(\phi, \theta) = 0 \iff \phi = (f_1(\theta), \ldots, f_q(\theta))$. De plus, le gradient de $f = (f_1, \ldots, f_q)$ est connu :

$$\nabla f = \left[\frac{\delta f_i}{\delta \theta_j}\right] = -J^{-1}H$$

avec J la matrice carrée $q \times q$ des $(\frac{\delta g_i}{\delta \phi_j})$ et H la matrice $p \times q$ des $(\frac{\delta g_i}{\delta \phi_j})$.

Ce résultat, déjà utilisé dans Benichou et Gail (1989), permet d'obtenir une estimation de la variance d'un estimateur d'un paramètre défini implicitement à partir des paramètres du modèle, et donc un intervalle de confiance approché de type Wald puisque la variance asymptotique de $\hat{\phi}_n$ peut être calculée et vaut $J^{-1}H\Sigma_n H' J^{-1'}$ si Σ_n est la matrice de variance asymptotique de $\hat{\theta}_n$ (les dérivées étant évaluées au point $\hat{\theta}_n$).

Si le paramètre d'intérêt est la racine $\phi \in \mathbb{R}$ d'une fonction réelle $g(., \theta)$, racine supposée unique, le résultat prend la forme plus simple

$$\hat{S}^{-1/2}(\hat{\phi} - \phi) \rightarrow_d \mathcal{N}(0, 1)$$

où

$$\hat{S} = \frac{1}{\left(\frac{\delta g}{\delta \phi}(\hat{\phi}_n, \hat{\theta}_n)\right)^2} \left[\frac{\delta g}{\delta \theta_1}, \ldots, \frac{\delta g}{\delta \theta_p}\right]_{|(\phi,\theta)=(\hat{\phi}_n,\hat{\theta}_n)} \Sigma_n \begin{bmatrix} \frac{\delta g}{\delta \theta_1} \\ . \\ . \\ . \\ \frac{\delta g}{\delta \theta_p} \end{bmatrix}_{|(\phi,\theta)=(\hat{\phi}_n,\hat{\theta}_n).} \quad (1)$$

Un intervalle de confiance approché au niveau γ peut donc être obtenu par

$$IC_\gamma(\phi) = \left[\hat{\phi}_n - z_{\frac{1+\gamma}{2}}\sqrt{S}; \hat{\phi}_n + z_{\frac{1+\gamma}{2}}\sqrt{S}\right]$$

.

1.2 Intervalles de confiance par bootstrap

Une méthode alternative repose sur le ré-échantillonnage par bootstrap, soit des données de l'échantillon, soit du paramètre estimé $\hat{\theta}$ connaissant une estimation de sa matrice de variance. Que le paramètre d'intérêt ϕ soit défini implicitement ne pose pas de difficulté de calcul des intervalles de confiance par bootstrap (voir Davison et Hinkley 1997, pour plus de détails).

2 Application au calcul de l'optimum altitudinal de la présence d'espèces végétales

Les changements climatiques de ces dernières années préoccupent les écologistes de par leur rapidité et leur amplitude. En effet ils auraient des conséquences plus ou moins importantes sur les niches écologiques des espèces végétales, notamment dans les domaines montagneux. De nombreuses études ont montré que l'analyse du gradient altitudinal d'une espèce permettrait de mettre en évidence l'influence du facteur climatique sur les écosystèmes au cours du temps notamment en comparant la valeur de l'altitude où l'espèce est la plus abondante pour deux périodes données. C'est dans le but d'évaluer l'adaptation des espèces végétales aux modifications du facteur climatique que le laboratoire Biogeco analyse des données de présence de cinq espèces de feuillus au sein de deux gradients altitudinaux situés dans deux chaines de montagnes en Espagne (les Pyrénées et le Système Ibérique) et pour deux inventaires forestiers séparés d'une période de 10 ans. Ainsi pour chaque espèce étudiée, l'altitude des plots à laquelle l'espèce était présente ou absente a été notée par inventaire et chaine de montagne. L'objectif est d'analyser la répartition d'une espèce le long d'un gradient de température, mais également d'étudier l'évolution de la répartition altitudinale de l'espèce au cours du temps suivant le déplacement de sa niche écologique engendré par exemple par une augmentation de température dans le cadre du réchauffement climatique.

Les données, utilisées pour l'étude, correspondent à celles répertoriées lors de deux inventaires forestiers espagnols, qui ont eu lieu respectivement entre 1986 et 1996, et, 1997 et 2007 (notés SFI1990 et SFI2000). Chaque SFI correspond à un échantillonnage d'arbres effectué selon une grille systématique de placettes permanentes à travers l'Espagne au sein desquelles a été relevée la présence ou non des différentes espèces d'arbres. L'ensemble de la surface forestière est ainsi échantillonnée sur une grille carrée de 1 km de côté. Chaque placette est localisée par ses coordonnées géographiques UMT. Au total, 73772 et 67542 placettes sont suivies lors du SFI1990 et SFI2000 respectivement. Pour chaque inventaire, on a sélectionné deux zones d'étude où ont été effectués les relevés : le système ibérique et les Pyrénées. Pour chaque chaîne montagneuse a été notée l'altitude exacte de chaque placette où ont été observée la présence ou non de plusieurs espèces d'arbres.

2.1 le modèle asymétrique HOF V

Pour une année et une zone d'étude d'un inventaire donnés, on dispose d'un échantillon $(X_i, Y_i)_{i=1...n}$ où i est l'indice de la placette, X_i son altitude et Y_i l'indicateur de présence de l'espèce d'arbre à analyser. Différentes modélisations existent dans la littérature, les deux principales reposant sur un modèle de régression logistique et un modèle de présence-absence plus général, tous les deux utilisant l'altitude seule comme variable explicative de la présence d'une espèce d'arbre.

Pour obtenir une courbe de réponse unimodale, il est courant (Oksanen et al, 2001) de considérer un modèle classique de régression logistique

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(a+bx+cx^2)}}, \tag{2}$$

les coefficients a, b et c faisant de la courbe de probabilité de présence une courbe concave présentant un optimum en $alt_{opt} = -\frac{b}{2c}$. La répartition de présence autour de l'optimum d'altitude présentant une symétrie qui peut être difficilement compatible avec les données, Huisman et al (1993) ont proposé un modèle concurrent qui présente l'avantage de supporter des courbes de répartition symétrique et assymétrique. Le modèle $HOF\ V$ est

$$P(Y = 1|X = x) = \frac{1}{1 + e^{a+bx}} \frac{1}{1 + e^{c+dx}} \tag{3}$$

où les coefficients a, b, c et d peuvent être contraints à prendre certaines valeurs fixées ce qui définit les sous modèles HOF IV (b=-d) et HOF III (d=0).

Les modèles HOF V et HOF IV (comme la régression logistique (2)) sont adaptés à la modélisation de la courbe de probabilité de présence des espèces végétales selon un gradient d'altitude car ils peuvent admettre un optimum en une valeur d'altitude noté alt_{opt}. Ceci-dit, seul le modèle HOF V présente une courbe de réponse asymétrique autour de son optimum mais, celui-ci n'étant pas explicite, le problème est celui du calcul de la variance de l'estimateur $\widehat{alt}_{opt}$ pour déterminer un intervalle de confiance.

La méthode d'estimation repose sur la maximisation de la vraisemblance, on dispose donc d'une estimation de la matrice de variance de l'estimateur $\hat{\theta}$.

2.2 Détermination d'un intervalle de confiance pour alt_{opt}

On considère le modèle HOF V, pour lequel la valeur de l'altitude optimale (notée ϕ dans la suite) est l'unique solution de

$$\frac{d}{dx}P(Y=1|X=x) \propto g(x,\theta) = b\ e^{a+bx} + d\ e^{c+dx} + (b+d)\ e^{(a+c)+(b+d)x} = 0$$

Pour appliquer les résultats de la section précédente, on calcule :

$$\begin{aligned}
\frac{\delta g}{\delta a}|_{(\hat{\phi},\hat{\theta})} &= -\hat{d}e^{\hat{c}+\hat{d}\hat{\phi}} \\
\frac{\delta g}{\delta b}|_{(\hat{\phi},\hat{\theta})} &= e^{\hat{a}+\hat{b}\hat{\phi}} + e^{\hat{a}+\hat{c}+(\hat{b}+\hat{d})\hat{\phi}} - \hat{d}\hat{\phi}e^{\hat{c}+\hat{d}\hat{\phi}} \\
\frac{\delta g}{\delta c}|_{(\hat{\phi},\hat{\theta})} &= -\hat{b}e^{\hat{a}+\hat{b}\hat{\phi}} \\
\frac{\delta g}{\delta d}|_{(\hat{\phi},\hat{\theta})} &= e^{\hat{c}+\hat{d}\hat{\phi}} + e^{\hat{a}+\hat{c}+(\hat{b}+\hat{d})\hat{\phi}} - \hat{b}\hat{\phi}e^{\hat{a}+\hat{b}\hat{\phi}}
\end{aligned}$$

et

$$\frac{\delta g}{\delta \phi}|_{(\hat{\phi},\hat{\theta})} = -\hat{b}\hat{d}\left[e^{\hat{a}+\hat{b}\hat{\phi}} + e^{\hat{c}+\hat{d}\hat{\phi}}\right]. \tag{4}$$

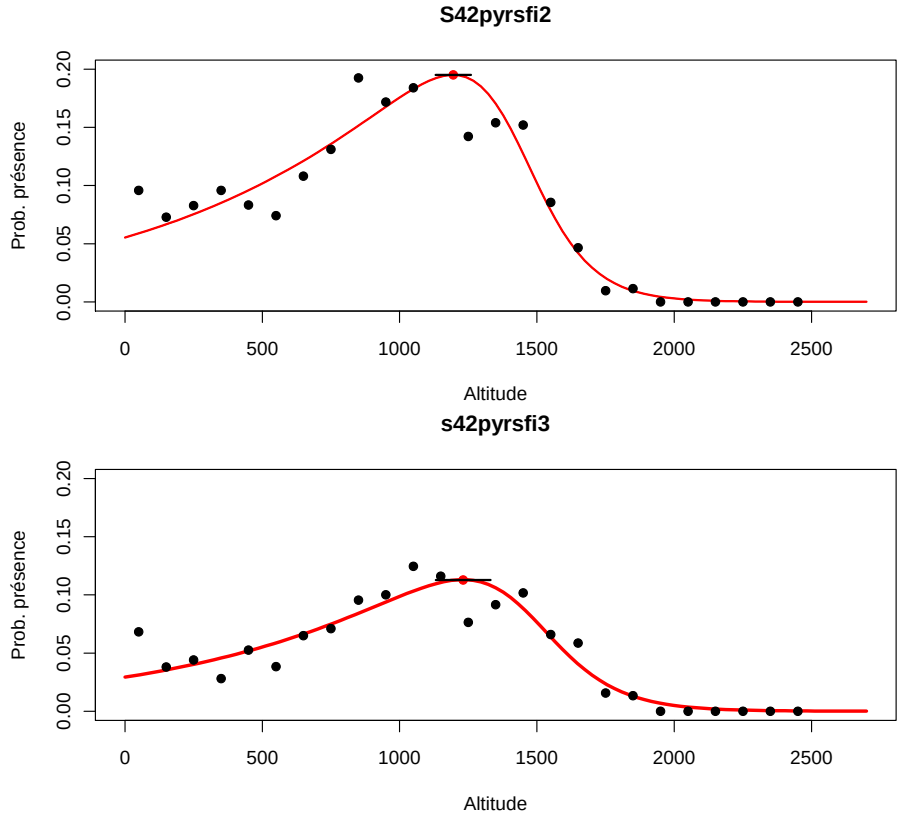
Ainsi, connaissant une estimation $\hat{\Sigma}$ de la matrice de variance de $\hat{\theta} = (\hat{a},\hat{b},\hat{c},\hat{d})$, l'application de (1) permet d'obtenir une estimation de la variance de $\widehat{alt_{opt}}$ et donc le calcul d'un intervalle de confiance approchée en utilisant la normalité approchée issue de la "méthode delta".

A titre d'exemple, on présente ci-dessous les résultats obtenus pour l'espèce *Quercus petraea* (chêne sessile) en 1995 et 2007. Les estimations sont résumées dans le tableau.

Période	a	b	c	d	alt_{opt}	$IC_{95\%}$.min	$IC_{95\%}$.max
SFI2 (1995)	2.837	-0.0013	-12.499	0.0087	1195	1131	1260
SFI3 (2007)	3.500	-0.0013	-11.008	0.0075	1231	1132	1331

Bibliographie

[1] Benichou, J. et Gail, M. (1989), A Delta Method for Implicitly Defined Random Variables, *The Amer. Statis.*, Vol. 43, No. 1 , pp. 41-44.
[2] Billingsley, P. (1986) *Probability and measure*, Wiley.



[3] Davison, A.C. et Hinkley, D.V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press.
[4] Huisman, J., Olff, H., Fresco, L.F.M. (1993) A hierarchical set of models for species response analysis, *Journal of Vegetation Science*, Vol 4, pp 37 - 46.
[5] Lawesson, J.E., Oksanen, J. (2002). Niche characteristics of Danish woody species as derived from coenoclines,*Journal of Vegetation Science*, Vol 13, pp 279 - 290.
[6] Lenoir, J., Gégout, J.C., Marquet, P.A., de Ruffray, P., Brisse, H. (2008) A significant upward shift in plant species optimum elevation during the 20th century, *Science*, Volume 320, p. 1768.
[7] Oksanen, J., Laara, E., Tolonen, K. et Warner, B. (2001), Confidence intervals for the optimum in the gaussian response function, *Ecology*, Vol 82, pp 1191-1197.
[8] Serfling, R.J. (1980) *Approximation Theorems of mathematical Statistics*, New York, John Wiley.
[9] Taylor, A.E. et Mann, W.R. (1983) *Advanced Calculus* (3rd ed.), New York, John Wiley.

Statistique des Processus

De nouvelles propriétés limites presque sûre pour les accroissements fonctionnels du processus empirique uniforme, *Davit Varron*

Comme l'a montré Wschebor en 1992, le mouvement brownien sur $[0,1]$ possède la propriété trajectorielle suivante : presque sûrement, lorsque u tend vers 0, la variable aléatoire $(W(t+u)-W(t))$ avec t distribué selon la mesure de Lebesgue- converge en loi vers une gaussienne centrée réduite. Nous montrons ici que cette propriété limite s'étend aux accroissements fonctionnels du processus empirique uniforme, et nous prouvons des résultats similaire concernant les loi fonctionnelles limites de type Strassen. L'outil principal de la preuve est un principe de Poissonisation dans a lignée de ceux connus jusqu'à lors, mais qui bénéficie d'une extension nouvelle.

Fast change point analysis on the Hurst index of piecewise fractional Brownian, *Mehdi Fhima, Pierre Bertrand and Arnaud Guillin*

In this presentation, we introduce a new method for change point analysis on the Hurst index for a piecewise fractional Brownian motion. We first set the model and the statistical problem. The proposed method is a transposition of the FDpV (Filtered Derivative with p-value) method introduced for the detection of change points on the mean in Bertrand et al. (2011) to the Hurst index. The underlying statistics of the FDpV technology is a new statistic estimator for Hurst index, so-called Increment Bernoulli Statistic (IBS). Both FDpV and IBS are methods with linear time and memory complexity, with respect to the size of the series.

Some mixing properties of conditionally independent processes, *Manel Kacem, Véronique Maume-Deschamps and Stéphane Loisel*

We consider conditionally independent processes, namely for all $i \in \mathbb{N}$, r.v's X_i are independent when conditioned to a factor $(V_1, ..., V_n)$. To our knowledge, conditionally independent variables with respect to a random vector with length time-varying have not been considered in the literature. It have some interest for modeling purposes. We study the mixing properties of these processes and derive self normalized central limit theorems. Some applications to risk processes in insurance are provided.

Théorèmes limites pour des martingales vectorielles à croissance explosive en temps continu et applications statistiques, *Hamdi Fathallah and Ahmed Kebaier*

Le but de l'exposé sera de présenter des résultats autour du théorème limite centrale presque-sûre pour des martingales vectorielles quasi-continues à gauche, en temps continu, et à croissance explosive. Je présenterai également une application des résultats obtenus à l'estimation du paramètre du modèle d'Ornstein-Uhlenbeck bivarié.

Test de comparaison de distributions pour des séquences fortement mélangeantes, *Laurence Reboul and Anne-Françoise Yao*

Dans ce travail, nous proposons un test lisse adaptatif de type Neyman pour la comparaison des distributions marginales de deux processus strictement stationnaires et mélangeants. Nous illustrons les performances de notre test au travers de simulations et applications à des données réelles.

De nouvelles propriétés limites presque sûres des accroissements fonctionnels du processus empirique uniforme

Davit VARRON

Laboratoire de Mathématiques Pures et Appliquées, UMR CNRS 6623, 16, route de Gray, 25030 Besançon Cedex

Abstract

Given an observation of the uniform empirical process α_n, its functional increments $\alpha_n(u+a_n\cdot)-\alpha_n(u)$ can be viewed as a single random process, when u is distributed under the Lebesgue measure. We investigate the almost sure limit behaviour of the multivariate versions of these processes as $n \to \infty$ and $a_n \downarrow 0$. Under mild conditions on a_n, a convergence in distribution and functional limit laws are established. The proofs rely on a new extension of usual poissonisation tools for the local empirical process.

Keywords : Empirical processes, limit theorems.

Résumé

Pour une réalisation particulière du processus empirique uniforme α_n (défini sur $(\Omega, \mathcal{A}, \mathbb{P})$), nous pouvons considérer ses incréments fonctionnels $\alpha_n(u + a_n\cdot) - \alpha_n(u)$ comme un processus stochastique, lorsque u est distribué selon la mesure de Lebesgue sur $[0, 1]$. Nous nous intéressons ici au comportement limite presque sûr (au sens de $(\Omega, \mathcal{A}, \mathbb{P})$), de ces "Lebesgue processus" (en considérant ici la version multivariée de α_n). Il est démontré une convergence en loi de processus, ainsi que deux loi fonctionnelles limites de type Strassen et non standard. Les preuves sont basée sur un nouveau lemme de poissonisation, qui fournit une nouvelle extension des lemmes déjà connus dans ce domaine.

Mots clés : Processus empiriques, théorèmes limites.

1 Introduction

En 1992, Wschebor [1] a découvert la propriété suivante pour le processus de Wiener $\mathbf{W}$ sur $[0, 1]$: presque sûrement, pour tous $0 \leq a < b \leq 1$, et pour tout borélien $B \subset \mathbb{R}$,

$$\lambda\Big(\big\{u \in [a, b],\ \epsilon^{-1/2}\big(\mathbf{W}(u+\epsilon) - \mathbf{W}(u)\big) \in B\big\}\Big) \underset{\epsilon\to 0}{\to} (b-a)\mathbb{P}\Big(\mathcal{N}(0,1) \in B\Big). \qquad (1)$$

Ici λ est la mesure de Lebesgue. Ce résultat fut ensuite étendu à une plus grande classe de processus par Azaïs et Wschebor [2]. Il est bien connu que les accroissements du processus empirique partagent un certain nombre de propriétés limite avec **W**. Une question qui vient alors naturellement à l'esprit est la suivant : est ce que ces incréments empiriques vérifient un propriété similaire à (1) ? Nous apportons un réponse positive à cette question, et même plus que cela :

- Une convergence similaire à (1) est établie, pour les incréments fonctionnels du processus empirique (théorème 1) ;
- Un équivalent de la loi limite de Strassen est démontré dans ce contexte particulier (théorème 2) ;
- Un équivalent de la loi limite fonctionnelle non standard est aussi démontré .
- Tous ces résultats sont valables pour le processus empirique uniforme basé sur un échantillon multivarié.

2 Résultats

Avant d'énoncer les résultats, nous devons introduire des notation. On notera $D([0,1]^d)$ l'espace de toutes les fonctions de répartition de mesures signées finies sur $[0,1]^d$ et $||\cdot||_{[0,1]^d}$ la norme sup sur $[0,1]^d$, c'est à dire :

$$|| f ||_{[0,1]^d} := \sup_{t\in[0,1]^d} | f(t) | .$$

Pour tous $f \in D([0,1]^d)$ et $A \subset [0,1]^d$ borélien, on notera $f(A)$ for $\mu(A)$, où μ est la mesure associée à f.
Soit $(U_n)_{n\geq 1}$ une suite i.i.d uniformément distribuée sur $[0,1]^d$. Pour $s = (s^{(1)}, \ldots, s^{(d)})$ et $t = (t^{(1)}, \ldots, t^{(d)})$ appartenant à $\mathbb{R}^d$ la notation $s \prec t$ sera sous entendue comme $s^{(k)} \leq t^{(k)}$ pour tous $k = 1, \ldots, d$. On notera aussi l'hypercube $[s,t] := [s^{(1)}, t^{(1)}] \times \ldots \times [s^{(d)}, t^{(d)}]$. Pour $u \in \mathbb{R}^d$ fixé et $a \in [0,1]$ on notera $u + a$ le vecteur $(u_1 + a, u_2 + a, \ldots, u_d + a)$ et on définit :

$$\Delta_n(u, a, \cdot) := \frac{\sum_{i=1}^{n} \left(\mathbb{1}_{[u,u+a\cdot]}(U_i) - \lambda([0, a\cdot]) \right)}{\sqrt{na^d}}.$$

On notera W la feuille brownienne standard (c'est à dire $\mathrm{Cov}(W(t), W(s)) := (s^{(1)} \wedge t^{(1)}) \times \ldots \times (s^{(d)} \wedge t^{(d)})$) et λ^* (resp. λ_*) la mesure Lebesgue extérieure (resp. intérieure) sur les sous ensembles de $[0,1]^d$.

Théorème 1 *Supposons que :*

$$a_n \downarrow 0, \quad na_n^d \uparrow \infty, \quad \liminf_{n\to\infty} \log(1/a_n)/\log\log(n) > 1. \tag{2}$$

Alors, presque sûrement, pour tout hypercube I vérifiant $\lambda(I) > 0$ et $I \subset [0, 1-\delta]^d$ pour un certain $\delta > 0$, les assertions suivantes sont vraies.

$$\begin{aligned}
&(i)\ \textit{Pour tout fermé} F \subset D([0,1]^d)\ \textit{on a} \\
&\quad \limsup_{n\to\infty} \frac{\lambda^*\big(\{u \in I,\ \Delta_n(u, a_n, \cdot) \in F\}\big)}{\lambda(I)} \le \mathbb{P}(W \in F), \\
&(ii)\ \textit{Pour tout ouvert} O \subset D([0,1]^d)\ \textit{on a} \\
&\quad \liminf_{n\to\infty} \frac{\lambda_*\big(\{u \in I,\ \Delta_n(u, a_n, \cdot) \in O\}\big)}{\lambda(I)} \ge \mathbb{P}(W \in O).
\end{aligned} \tag{3}$$

Notre second résultat est un résultat similaire 1, mais dans l'esprit des lois de Strassen. On notera J la fonction de taux des grandes déviations de W, c'est à dire :

$$J(f) := \inf\left\{ \int_{[0,1]^d} g^2(u)du,\ f = \int_{[0,\cdot]} g(s)ds \right\},\ f \in D([0,1]^d), \tag{4}$$

avec la convention $\inf_\emptyset = +\infty$. La définition précédente nous permet de définir la boule de Strassen par

$$\mathcal{S} := \Big\{ f \in D([0,1]^d),\ J(f) \le 1 \Big\}. \tag{5}$$

Théorème 2 *Supposons que*

$$a_n \downarrow 0,\ \ na_n^d \uparrow \infty,\ \ \frac{na_n^d}{\log\log(n)} \to \infty,\ \ \liminf_{n\to\infty} \frac{\log(1/a_n)}{\log\log(n)} > 2. \tag{6}$$

Alors, presque sûrement, pour tout hypercube I vérifiant $\lambda(I) > 0$ et $I \subset [0, 1-\delta]^d$ pour un certain $\delta > 0$, on a

$$\frac{\lambda\left(\left\{u \in I,\ \frac{\Delta_n(u,a_n,\cdot)}{\sqrt{2\log\log(n)}} \rightsquigarrow \mathcal{S}\right\}\right)}{\lambda(I)} = 1. \tag{7}$$

Ici $f_n \rightsquigarrow \mathcal{S}$ signifie que la suite $(x_n)_{n\ge1}$ admet $\mathcal{S}$ pour ensemble d'adhérence dans le Banach $D([0,1]^d)$.

Notre troisième résultat est dans la lignée des lois fonctionnelles limites non standard, apparaissant lorsque $na_n^d \sim c\log\log(n)$ avec $0 < c < \infty$. Nous devons d'abord définir la fonction de taux associée aux grandes déviations d'un processus de Poisson sur $\mathbb{R}^d$.

$$\mathfrak{J}(f) := \inf\left\{ \int_{[0,1]^d} h(u)du,\ f = \int_{[0,\cdot]} g(s)ds \right\},\ f \in D([0,1]^d), \tag{8}$$

avec $h(x) := x\log(x) - x + 1$ for $x > 0$ et $h(0) := 0$. Pour $c > 0$ on notera

$$\Gamma_c := \Big\{ f \in D\big([0,1]^d\big),\ \mathfrak{J}(f) \le 1/c \Big\}.$$

Théorème 3 *Supposons que* $na_n^d \sim c\log\log(n)$ *pour un certain* $0 < c < \infty$. *Alors, presque sûrement, pour tout hypercube* I *vérifiant* $\lambda(I) > 0$ *et* $I \subset [0, 1-\delta]^d$ *pour un certain* $\delta > 0$, *on a*

$$\frac{\lambda\left(\left\{u \in I,\ \frac{\Delta F_n(u,a_n,\cdot)}{c\log\log(n)} \rightsquigarrow \Gamma_c\right\}\right)}{\lambda(I)} = 1, \tag{9}$$

où

$$\Delta F_n(u, a_n, t) := \sum_{i=1}^{n} \mathbb{1}_{[u,u+a_n t]}(U_i),\ u, t \in [0,1]^d. \tag{10}$$

Dans chacune des trois preuves nous faisons systématiquement appel à

- Un outil permettant de remplacer les incréments du processus empirique par leur versions *poissonisées*. Ces versions poissonisées ont des propriétés (indépendance des "incréments") permettant de mener simplement des calculs de variance.
- La connaissance existante sur le comportement de ces versions poissonisées des $\Delta_n(0, a_n, \cdot)$).

Bibliographie

[1] Wschebor, M.(1992) *Sur les accroissements du processus de Wiener*, C.R. Acad. Sci. Paris, Ser. I, 315(12) :1293–1296.

[2] Azaïs, J.M. et Wschebor, M (1996) *Almost Sure Oscillation of Certain Random Processes*, Bernoulli, 2(3) :257–270.

Fast change point analysis on the Hurst index of piecewise fractional Brownian

Pierre, R. BERTRAND[1,2] *Pierre.Bertrand@math.univ-bpclermont.fr*

Mehdi FHIMA[2] *Mehdi.Fhima@math.univ-bpclermont.fr*

Arnaud GUILLIN[2] *Arnaud.Guillin@math.univ-bpclermont.fr*

[1] *INRIA Saclay*

[2] *Laboratoire de Mathématiques, UMR CNRS 6620*
& Université de Clermont-Ferrand II, France

Résumé: Dans cette présentation, nous introduisons une nouvelle méthode de détection de ruptures sur l'indice de Hurst pour un mouvement brownien fractionnaire par morceaux. En premier, nous définissons le modèle et le problème statistique. La méthode proposée est une transposition de la FDpV à l'indice de Hurst. La FDpV (dérivée filtrée avec p-valeur) est une méthode introduite pour détecter des ruptures sur la moyenne par Bertrand et al. (2011). La statistique sous-jacentes de la technologie FDpV est un nouvel estimateur de l'indice de Hurst, appelé Increment Bernoulli statistique (IBS). A la fois les méthodes FDpV et IBS ont une complexité linéaire par rapport à la taille de la série d'observation, aussi bien en temps de calcul que pour la mémoire.
Mots clés: Détection de ruptures, Dérivée Filtrée, mouvement Brownien fractionnaire par morceaux, paramètre de Hurst, Increment Bernoulli statistique.

Abstract: In this presentation, we introduce a new method for change point analysis on the Hurst index for a piecewise fractional Brownian motion. We first set the model and the statistical problem. The proposed method is a transposition of the FDpV (Filtered Derivative with p-value) method introduced for the detection of change points on the mean in Bertrand et al. (2011) to the Hurst index. The underlying statistics of the FDpV technology is a new statistic estimator for Hurst index, so-called Increment Bernoulli Statistic (IBS). Both FDpV and IBS are methods with linear time and memory complexity, with respect to the size of the series.
Keywords: Change point analysis, Filtered derivative with p-value method, Hurst parameter, Increment Bernoulli Statistic, piecewise fractional Brownian motion.

Introduction

Recent measurement methods allow us to record and to stock large data sets, so called "the data deluge". For instance, today technology allows recording of heartbeat series during 24 hours leading to data sets of size $n \geq 100,000$, and very high frequency (VHF) financial series leads to data size $n \geq 40,000$. Tomorrow, many other series will be recorded at VHF leading to millions of data.

Large or huge series with small meshes of time can be described as continuous time processes observed at discrete times. Such a stochastic process X belongs to a certain class of model, that is $X \in \mathcal{M} = \{X_\theta,\ \theta \in \Theta\}$, where Θ is a subset of $\mathbb{R}^d$ and d is the dimension of the model. The structural parameter θ is believed to provide relevant information on the system which generate the series, and statisticians have to estimate it.

A slightly different approach is based on change point analysis: The structural parameter θ is assumed to be piecewise constant with an unknown configuration of change τ. In this framework, the first task of statisticians is the estimation of the location of the change points and the second one could be the estimation of the structural parameters between change points. There is a huge literature on change point analysis and model selection since the fifties, see e.g. Basseville & Nikiforov (1993), Brodsky and Darkhovsky (1993), Csörgo and Horváth (1997), Birgé and Massart (2007) or Bertrand *et al* (2011) and the references therein. However, most of the studies are devoted to change on the mean, on the variance or on the regression parameters. But relevant informations are also provided by the time dependence structure of the process, see e.g. Ayache & Bertrand (2011) and Khalfa *et al* (2011). Fractional Brownian motion (fBm) is a paradigmatic example of such process, indeed fBm is a zero mean Gaussion process depending on two parameters: The Hurst index H linked to the time structure and a scale parameter σ.

In this presentation, we consider a simple model, that is a process X which is a piecewise fBm with an unknown configuration of changes. Moreover, we set us in the frame of huge datasets, and we focus our attention on time and memory complexity. These two reasons have lead us to propose a new change point procedure for detection of change on the Hurst index for a piecewise fBm. Our new procedure is the combination, on the one hand of the FDpV method, introduced in Bertrand *et al* (2011) for fast and light detection of change on the mean, variance or regression parameter, and on the other hand of the Increment Bernoulli Statistic (IBS) a new estimator for Hurst index, which is a variation on the Increment Ratio Statistic (IRS) estimator introduced in Bardet and Surgailis (2010).

The rest of this paper is organized as follows: At first, in Section 1, we define our model of piecewise fBm. Next, in Section 2, we introduce a new fast and robust estimator of the Hurst index of fBm, namely the Increment Bernoulli Statistic. Then, in Section 3,we describe the transposition of the FDp-V method to Hurst index.

1 Our model

We observe a process X at the discrete and regularly spaced time $t_i = i/n$, where $i = 0, \ldots, n$. We assume the existence of a segmentation $\tau = (\tau_k)_{k=0,\ldots,K+1}$, with $0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n$, such that the restriction of the process X on each interval $(\tau_k/n, \tau_{k+1}/n)$ for $k = 0, \ldots, K$ is a fBm with Hurst index H_k and scale parameter σ_k. The integer K corresponds to the number of change points and $(K + 1)$ to the number of segments. Stress that K can be zero and in this case the process X is a fBm.

Let us precise that in our roadmap the process X should be almost surely with continuous paths. For this reason, the so-called piecewise fBm can not be defined by plugging a piecewise Hurst index into one of the representations of the fBm. Indeed, by doing so, the process X would be almost surely discontinuous at change point on Hurst index, and method for detecting change on the mean will also be efficient for detecting change on Hurst index. Let us refer to Taqqu and Samorodinitsky (1994) as a reference book on the different representation of fBm and to Ayache and Taqqu (2005) for the construction of multi-fBm by plugging a continuous time varying Hurst index into one of the fBm representations. A rather complicated solution to avoid this drawback was proposed in Benassi *et al* (2000) and cosigned by the first author. However, as point out by

Antoine Ayache during private conversations held in 2004, for statistical applications, it suffices to cancel the discontinuity by adding a correction term. This solution is also adopted in Bardet and Kammoun (2008) .

The model having being specified, we are concerned with change point analysis on the Hurst parameter, where the number of change K is unknown. There are few references on this problem. To our best knowledge, the only reference are Benassi *et al* (2001) and Bardet & Kammoun (2008).

2 Increment Bernoulli Statistic for fBm

In this section, we investigate the properties of a new estimator of the Hurst parameter of fBm, namely the Increment Bernoulli Statistic (IBS). IBS is a variation on IRS which has been introduced by Surgailis *et al* (2008) and applied to fBm by Bardet and Surgailis (2010). Both IRS and IBS are fast and robust estimator of the Hurst index. By fast we mean estimator with linear time complexity, and by robust we mean estimator with invariant scaling property. The choice of the IBS instead of the IRS is motivated by the fact that the IBS is bit less expensive, in terms of time complexity, than the IRS.

In the next section, the IBS is used as the underlying estimator for the FDpV method, see (1). For this reason, we define IBS for a every process X, even if we apply it to fBm in this section. Let X be a process observed at a family of discrete times t_k, we define the second order increments by

$$\Delta X(t_k) := X(t_{k+2}) - 2X(t_{k+1}) + X(t_k).$$

Then, the Increment Bernoulli Statistic (IBS) is based on the comparison of the signs of consecutive second order increments. The results of these comparisons will be equal to 1 if the consecutive second order increments have the same sign, and 0 otherwise. Hence, this explains the name of our new estimator, that is to say: Increment Bernoulli Statistic (IBS) which is given by

$$\mathrm{IBS}_n(X) = \frac{1}{n-2} \sum_{k=0}^{n-3} \psi\left(\Delta X(t_k), \Delta X(t_{k+1})\right)$$

where $\psi(\cdot,\cdot)$ is described as follows $\psi(x,y) = 0$ if $\mathrm{sign}(x) = \mathrm{sign}(y)$ and 0 otherwise, where $\mathrm{sign}(z)$ denotes the sign of z. Let us remark that IBS is scale invariant: Indeed, since $\psi(\sigma x, \sigma y) = \psi(x,y)$ for $\sigma > 0$, then $\mathrm{IBS}_n(\sigma X) = \mathrm{IBS}_n(X)$.

When X is a fBm, that is $X = B_H$ with a Hurst index $H \in (0,1)$, then the IBS converges in distribution to a continuous monotonic increasing function $\Lambda(H)$ defined as follows

$$\begin{aligned} \Lambda(H) &:= \Lambda_0\left(\rho(H)\right) \\ \Lambda_0(r) &:= \frac{1}{\pi}\arccos(-r) \\ \rho(H) &= \left(-3^{2H} + 2^{2H+2} - 7\right)\left(8 - 2^{2H+1}\right)^{-1} \end{aligned}$$

where $\rho(H) \in (-1,1)$ represents the correlation between two successive second order increments. The graph of $\Lambda(H)$ is given in Figure 1. Then, due to the fact that $\Lambda(\cdot)$ is a reversible function,

it is easy to deduce an estimator of the Hurst parameter H given by $\widehat{H}_n = \Lambda^{-1}(\mathrm{IBS}_n(B_H))$. Furthermore, we note that the function $\phi(\cdot,\cdot) = \psi(\cdot,\cdot) - \Lambda(H)$ is a Hermite function with rank equal to 2. Then, by applying the Breuer-Major theorem, see e.g Arcones (1994)[Theorem 4, p.2256] or Nourdin *et al* (2010)[Theorem 1, p.2], we can deduce the following Central Limit Theorem (CLT):

$$\sqrt{n}\left(\mathrm{IBS}_n(B_H) - \Lambda(H)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(H)),$$

where the sign $\xrightarrow{\mathcal{D}}$ means convergence in distribution and the asymptotic variance $\sigma^2(H)$ is given by

$$\sigma^2(H) = \sum_{j\in\mathbb{Z}} \mathrm{cov}\left(\psi\left(\Delta B_H(t_0), \Delta B_H(t_1)\right), \psi\left(\Delta B_H(t_j), \Delta B_H(t_{j+1})\right)\right).$$

The main advantages of the IBS method are primarily its efficiency in terms of time and memory complexity, and secondarily its robustness against scaling properties of the fBm. At first, we calculate by recurrence the second order increments $(\Delta B_H(t_k))_{0\leq k\leq n-3}$, this step is performed in time and memory complexity on $\mathcal{O}(n)$. Next, the computing of $\mathrm{IBS}_n(B_H)$ requires roughly n tests $+$ $n \times p_a(H)$ additions $+$ 1 division, where $p_a(H) = \Lambda(H) \in (0,1)$ corresponds to the probability that two consecutive second order increments have the same sign. Then, we apply the Newton algorithm to compute the inverse of the function $\Lambda(\cdot)$. Moreover, we note that the function $\psi(\cdot,\cdot)$ satisfy the scale invariant property, *i.e.* for all $C \in \mathbb{R}$, $\psi(C.X, C.Y) = \psi(X,Y)$. This means that the multiplication of B_H by any scaling coefficient C does not have any effect on the estimation of the Hurst index since $\mathrm{IBS}_n(B_H) = \mathrm{IBS}_n(C.B_H)$. Hence, this proves the robustness of the IBS method against scaling properties of the fBm.

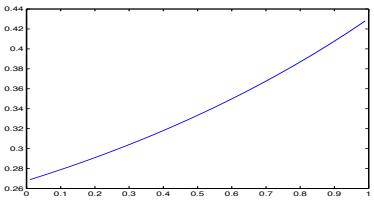


Figure 1: *The graph of* $\Lambda(H)$.

3 Filtered Derivative with p-Value method

In this section, we describe the Filtered Derivative with p-value method (FDp-V). First, we define the Filtered Derivative function. Next, we describe the two steps of the FDp-V method: Step 1 is based on Filtered Derivative and select the potential change points, whereas Step 2 calculate the p-value associated to each potential change point, for disentangling right change points and false alarms.

At first, we note that $\Lambda(H)$ is a continuous monotonic increasing function of H, see Figure 1. So, the detecting of change points on the Hurst parameter H is equivalent to detecting change points on $\Lambda(H)$. Consequently, the estimator $\mathrm{IBS}_n(B_H)$ of the parameter $\Lambda(H)$ is used as the underlying estimator for the FDpV method. We reefer to Bertrand *et al* (2011) and Bertrand & Fhima (2009) for the introduction of FDpV technology and its numerical efficiency. Let us stress that the choice of the direct estimator $\widehat{H}_n = \Lambda^{-1}(\mathrm{IBS}_n(B_H))$ as the underlying estimator for the FDpV would be more expensive in term of numerical complexity.

Filtered Derivative function

Let X be a piecewise fBm observed at a family of discrete times $t_j = j/n$, for $j = 0, \ldots, n$. The Filtered Derivative for IBS is defined as the difference between the estimators of the parameter $\Lambda(H)$ computed on two sliding windows respectively at the right and at the left of the index k, both of size A, that is specified by the following function

$$D(k, A) = \mathrm{IBS}\,(X, k, A) - \mathrm{IBS}\,(X, k - A, A) \text{ for } k \in [A, n - A] \tag{1}$$

where

$$\mathrm{IBS}\,(X, k, A) = A^{-1} \sum_{j=k+1}^{k+A} \psi\left(\Delta X(t_k), \Delta X(t_{k+1})\right)$$

is an estimator of $\Lambda(H)$ on the sliding box $[k + 1,\, k + A]$. It is easy to see that the Filtered Derivative function D is computed by recurrence with linear time and memory complexity. Eventually, this method consists on filtering data by computing the estimators of the parameter $\Lambda(H)$ before applying a discrete derivation. This construction explains the name of the algorithm, so-called Filtered Derivative method, by Benveniste and Basseville (1984).

Step 1: Detection of potential change points

In order to detect the potential change points, we test the null hypothesis $(\mathcal{H}_0)$ of no change in the Hurst parameter H against the alternative hypothesis $(\mathcal{H}_1)$ indicating the existence of at least one change point

$(\mathcal{H}_1)$: There is an integer $K \in \mathbb{N}^*$ and $0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n$ such that

$$H_1 = \cdots = H_{\tau_1} \neq H_{\tau_1+1} = \cdots = H_{\tau_2} \cdots \neq H_{\tau_K+1} = \cdots = H_{\tau_{K+1}}.$$

where $H_j = H_{\tau_k}$ is the value of the Hurst parameter at $t_j \in [\tau_{k-1}/n, \tau_k/n)$.
Now, we fix a probability of type I error at level p_1^*, and we determine the corresponding critical value C_1 given by

$$\mathbb{P}\left(\max_{k\in[A:n-A]} |D(k, A)| > C_1 | \mathcal{H}_0 \text{ is true}\right) = p_1^*.$$

Of course, such a probability is usually not available, so that we only consider the asymptotic distribution of the maximum of $|D|$. Then, the change points $\widetilde{\tau}_k$ is selected as a potential change point if its local maxima satisfy $|D(\widetilde{\tau}_k, A)| > C_1$. We remark through the graph of the function $|D|$ that there are not only the "right hats" (surrounded in green in Figure 2) which gives the right change points, but also false alarms (surrounded in black in Figure 2). Consequently, we have introduced another step in order to keep just the right change points.

Step 2: Elimination of false alarms

The list of potential change points $(\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{K_{\max}})$ obtained at step 1 contains right change points but also false detections. In the second step a test is carried out to remove the false alarms from the list of change points found at step 1. More precisely, for all potential change point $\widetilde{\tau}_k$, we test whether the Hurst parameter is the same on the two successive intervals $(\widetilde{\tau}_{k-1}/n, \widetilde{\tau}_k/n)$ and $(\widetilde{\tau}_k/n, \widetilde{\tau}_{k+1}/n)$, or not. Formally, for all $1 \leq k \leq K_{\max}$, we apply the following hypothesis testing

$$(\mathcal{H}_{0,k}) : H_k = H_{k+1} \quad \text{versus} \quad (\mathcal{H}_{1,k}) : H_k \neq H_{k+1},$$

where H_k is the value of H on the segment $(\widetilde{\tau}_{k-1}/n, \widetilde{\tau}_k/n)$. By using this second test, we calculate new p-values $(\widetilde{p}_1, \ldots, \widetilde{p}_{K_{\max}})$ associated respectively to each potential change points $(\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{K_{\max}})$. Then, we only keep the change points which have a p-value smaller than a critical level denoted p_2^*. By doing so, we obtain a subset $\left(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}}\right)$ of the first list which represents the estimators of the change points in the Hurst parameter of mBm.

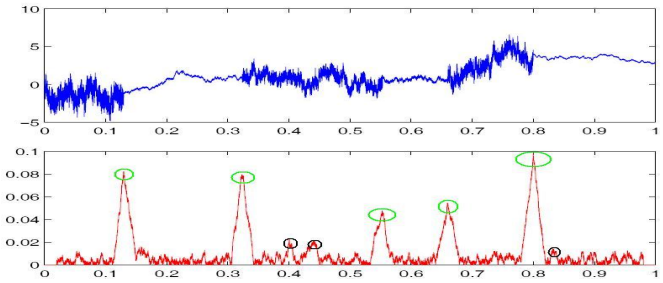


Figure 2: *Detection of potential change points. Above: Simulated piecewise fBm with five change points in the Hurst parameter. Below: Filtered Derivative function |D|.*

Conclusion

In conclusion, it appears that the combination of the FDpV and the IBS methods provides a fast (time) and cheap (memory) algorithm to the detection of change points on the Hurst parameter of piecewise fBm. So, this algorithm is adapted to segment random signals with large datasets. In future work, we will develop the FDpV + IBS method in order to detect abrupt changes on parameters of real data drawn from financial and physiological domains.

References

[1] Antoch, J. and Hušková, M. (1994) Procedures for the detection of multiple changes in series of independent observations, *Asymptotic statistics (Prague, 1993)*, Contrib. Statist., 3–20. Physica, Heidelberg.

[2] Arcones, M. A (1994) Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors, *Ann. Probab*, 22, 2242–2274.

[3] Ayache, A. and Bertrand, P.R. (2011) Discretization error of wavelet coefficient for fractal like process, *To appear in Advances in Pure and Applied Mathematics.*
[4] Ayache, A. and Taqqu, M. S. (2005) Multifractional process with random exponent, *Publ. Mat*, 49, 459–486.
[5] Bardet, J. M. and Kammoun, I. (2008). Detecting abrupt changes of the long-range dependence or the self-similarity of a Gaussian process, *C. R. Math. Acad. Sci.* Paris, 346, 889–894.
[6] Bardet, J. M. and Surgailis, D. (2010) Measuring Roughness of Random Paths by Increment Ratios, *To appear in Bernoulli.*
[7] Basseville, M. and Nikiforov, I. V. (1993) *Detection of abrupt changes: theory and application*, Prentice Hall Inc, Englewood Cliffs, NJ.
[8] Benassi, A., Bertrand, P. R., Cohen, S., Istas, J. (2000) Identification of the Hurst exponent of a Step Multifractional Brownian motion, *Statistical Inference for Stochastic Processes*, 3, 101–111.
[9] Benveniste, A. and Basseville, M. (1984) Detection of abrupt changes in signals and dynamical systems: some statistical aspects, *Analysis and optimization of systems, Part 1 (Nice, 1984), volume 62 of Lecture Notes in Control and Inform. Sci.*, 145–155, Springer, Berlin.
[10] Bertrand, P. R. and M. Fhima, M. (2009) Filtered Derivative with p-Value Method for Multiple Change-Points Detection, *Proceeding of the 2nd International Workshop in Sequential Methodologies.*
[11] Bertrand, P. R., Fhima, M. and Guillin, A. (2011) Off-line detection of multiple change points by the Filtered Derivative with p-Value method, *To appear Sequential Analysis.*
[12] Birgé, L. and Massart, P. (2007) Minimal penalties for Gaussian model selection, *Probab. Theory Related Fields*, 138(1-2), 33–73.
[13] Brodsky, B. E. and Darkhovsky, B. S. (1993) *Nonparametric methods in change-point problems, volume 243 of Mathematics and its applications*, Kluwer Academic Publishers Group, Dordrecht.
[14] Csörgo, M. and Horváth, L. (1997) *Limit Theorem in Change-Point Analysis*, J. Wiley, New York.
[15] Khalfa, N., Bertrand, P. R., Boudet, G., Chamoux, A. and Billat, V. (2011), Heart Rate Regulation processed through wavelet analysis and change detection. Some case studies, *Submitted.*
[16] Nourdin, I., Peccati, G. and Podolskij, M. (2010) Quantitative Breuer-Major Theorems, HAL : hal-00484096, version 2.
[17] Samorodnitsky, G. and Taqqu, M. S. (1994) *Stable non-Gaussian random processes*, Chapman & Hall.
[18] Surgailis, D., Teyssière, G. and Vaičiulis, M. (2008) The increment ratio statistic, *J. Multivariate Anal*, 99, 510–541.

Some mixing properties of conditionally independent processes

Kacem Manel, Loisel Stéphane & Maume-Deschamps Véronique

Univérsité de Lyon, F-69622, Lyon, France, Université Claude Bernard Lyon 1 Laboratoire SAF, EA 2429, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, F-69007 Lyon, France

Abstract

We consider conditionally independent processes, namely for all $i \in \mathbb{N}$, r.v's X_i are independent when conditioned to a factor $(V_1, \ldots, V_n)$. To our knowledge, conditionally independent variables with respect to a random vector with length time-varying have not been considered in the literature. It have some interest for modeling purposes. We study the mixing properties of these processes and derive self normalized central limit theorems. Some applications to risk processes in insurance are provided.

Résumé

On considère un processus conditionnellement indépendant tel que pour tout $i \in \mathbb{N}$, les variables aléatoires X_i sont indépendantes en conditionnant par rapport à un facteur $(V_1, \ldots, V_n)$. A notre connaissance, les variables conditionnellement indépendantes par rapport à un vecteur aléatoire de longueur variante au cours du temps, qui représente la mémoire du facteur, n'étaient pas considérées en littérature. Pour des fins de modélisation il est intéressant de les examiner. Dans ce cadre en notant par $\underline{V}_j = (V_1, \ldots, V_j)$ la mémoire du facteur pour tout $j \in \mathbb{N}$ et par $N^j(dx_j, \underline{V}_j)$ le noyau de transition de X_j sachant $\underline{V}_j$, on considère que $\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u})|\underline{V}_n) = \int_{x_{i_1}, \ldots, x_{i_u}} f(x_{i_1}, \ldots, x_{i_u}) \prod_{j=i_1}^{i_u} N^j(dx_j, \underline{V}_j)$. Nous étudions les propriétés de mélange de ces processus. Sous une condition supplémentaire de mélange fort conditionnel nous prouvons l'asymptotique normalité de S_n avec $S_n = \sum_{i=1}^n X_i$. Nous proposons un estimateur pour la variance et nous obtenons un théorème central limite normalisé. Nous présentons quelques applications aux processus de risque en assurance.

Keywords: Conditional independence; Common-factor; Controlling random vector with varying length; Mixing; Conditional mixing; Central Limit Theorem; Asymptotic stationarity; Variance estimation.

Mots clés : Indépendance conditionnelle ; Facteur commun ; Vecteur aléatoire de conditionnement de longueur variante ; Mélange ; Mélange conditionnel ; Theorem Central Limite ; Stationnarité asymptotique ; Estimation de la variance.

Acknowledgements. This work has been supported by the French Research National Agency (ANR) under refrence ANR-08-BLAN-0314-01.

1 Introduction

In our model we assume that dependence among risks may be explained by the long memory of a common factor. This common factor may influence the aggregate claim amount of a risk or simply the occurrence of a risk. Note that in actuarial science such a model is considered with common chock in which dependence between risks may be derived by a varying-time-common-factor. This factor can modulate the economic environment, the state of nature or the state of health in a population. As an example Cossette and Marceau (1999) studied a Poisson risk model with common chock represented by a discrete random variable. More recently, Cossette and al (2004) proposed a compound binomial model modulated by a markovian environment and they provided algorithms to compute the aggregate claim amount distribution in a fixed time period and the numerical values of infinite-time ruin probabilities. Denote the factor by the random variable (r.v) V and $\underline{V}_n$ by $(V_1, \ldots, V_n)$. Cossette and al (2004) considered that for all $i \in \mathbb{N}$, r.v's X_i are conditionally independent given r.v's V_i, roughly speaking $\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u})|\underline{V}_n) = \int_{x_{i_1},\ldots,x_{i_u}} f(x_{i_1}, \ldots, x_{i_u}) \prod_{j=i_1}^{i_u} N^j(dx_j, V_j)$ where $N^j(dx_j, V_j)$ denote the Kernel transition of X_j given V_j. In our framework, r.v.'s $(X_n)_{n\in\mathbb{N}}$ are considered to be conditionally independent given the entire trajectory of the factor that is $\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u})|\underline{V}_n) = \int_{x_{i_1},\ldots,x_{i_u}} f(x_{i_1}, \ldots, x_{i_u}) \prod_{j=i_1}^{i_u} N^j(dx_j, \underline{V}_j)$, where $N^j(dx_j, \underline{V}_j)$ denote the Kernel transition of X_j given $\underline{V}_j = (V_1, \ldots, V_j)$. We note that in this last case sequences $(\underline{V}_j)_{(j\in\mathbb{N})}$ overlapped. Denote by $S_n = \sum_{j=1}^n X_j$. Under additional condition of conditional strong mixing (see definition 1) we prove asymptotic normality of the sum S_n. In order to make Central Limit Theorem applicable, we provide a consistent estimator of the variance of the sum adapted with our mixing coefficients. Note that in our model $(V_n)_{n\in\mathbb{N}}$ is not necessarily a discrete random sequence, it can be a continuous time series. We give a simple example to illustrate our framework. Other applications related to more complex real insurance model are in work progress.

2 Mixing coefficients, conditional independence and conditional strong mixing condition

We define Mixing sequences of random variables as sequences for which past and future are asymptotically independent. Let u, v be positive integers, a random process $(X_1, \ldots, X_n)$ is said to be $\eta(u, v)$-mixing if for any real valued bounded functions f and g, for any integers $u' \leq u$, $v' \leq v$ and r and for any multi-index $i_1 \leq \cdots \leq i_{u'} \leq i_u < i_u + r \leq j_1 < \cdots \leq j_{v'} \leq j_v$,

$$sup \left|\text{Cov}\left(f(X_{i_1}, \ldots, X_{i_{u'}}), g(X_{j_1}, \ldots, X_{j_{v'}})\right)\right| \leq C(u, v)\eta(r)\|f\|_a\|g\|_b, \qquad (2.1)$$

where the supremum is taken over all the sequences $(i_1, \ldots, j_v)$, $r = j_1 - i_u$ is the gap of time between past and future and $\eta(r) \overset{r\to\infty}{\longrightarrow} 0$. $\| \ \|_a$ and $\| \ \|_b$ are norms, with respect

to this norm, we have various kind of mixing. If (2.1) is satisfied for all u, uniformly in u (that this $C(u,v) = C(v)$), we shall say that the process is $\eta(\infty, v)$ mixing. If (2.1) is satisfied for all v, uniformly in v (that this $C(u,v) = C(u)$), we shall say that the process is $\eta(u, \infty)$ mixing. If (2.1) is satisfied for all u and v, uniformly in u and v (that this $C(u,v) = C$), we shall say that the process is η mixing. According to the choice of $\| \ \|_a$ and $\| \ \|_b$, we recover the classical definitions of α, Φ, Ψ mixing. Peligrad (1986) and Bradley (2005) provided a unified presentation of some of the principal classical mixing coefficients and their properties. Note that for the classical mixing coefficients, $C(u,v)$ is uniform in u and v.

Remark 1 *For other choices of $\| \ \|_a$, $\| \ \|_b$ (for example Lipschitz norms or bounded variation norms), we shall recover the notions of weak dependance (see Dedecker and al (2007), Dedecker and Prieur(2005), Dedecker and Doukhan(2003) and Doukhan and Louhichi(1999)).*

2.1 Conditional independence

Assume that

- H1: For all $i \in \mathbb{N}$, r.v's X_i are conditionally independent given $\underline{V}_i = (V_1, \ldots, V_i)$.
- H2: Conditional law of $X_i|\underline{V}_n$ is the same of $X_i|\underline{V}_i$, this is for all $n \geq i$ and $(n,i) \in \mathbb{N}^2$.
  For all u-tuple, $(i_1, \ldots, i_u)$ such that $i_1 \leq i_2 \ldots \leq i_u$ these two assumptions mean that for all $n > i_u$

$$\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u}|\underline{V}_n)) = \int_{x_{i_1}, \ldots, x_{i_u}} f(x_{i_1}, \ldots, x_{i_u}) \prod_{j=i_1}^{i_u} N^j(dx_j, \underline{V}_j), \tag{2.2}$$

  Where $N^j(dx_j, \underline{V}_j)$ denote the Kernel transition of X_j given $\underline{V}_j$.

Proposition 1 *Let the v-tuple, $(j_1, \ldots, j_v)$ and the u-tuple, $(i_1, \ldots, i_u)$ be such that $i_1 \leq i_2 \ldots \leq i_u < i_u + r \leq j_1 \leq j_2 \ldots \leq j_v$ and let f and g be real valued bounded functions. If $H1$ and $H2$ hold then we obtain*

$$Cov(f(X_{i_1}, \ldots, X_{i_u}), g(X_{j_1}, \ldots, X_{j_v})) = Cov(\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u})|\underline{V}_{i_u}), \mathbb{E}(g(X_{j_1}, \ldots, X_{j_v})|\underline{V}_{j_v})), \tag{2.3}$$

2.2 Conditionally strongly mixing sequences

To introduce our dependence framework we define conditionally strongly mixing sequences as the following

Definition 1 *Let the u-tuple, $(i_1, \ldots, i_u)$ and the v-tuple, $(j_1, \ldots, j_v)$ be such that $i_1 \leq i_2 \ldots \leq i_u < i_u + r \leq j_1 \leq j_2 \ldots \leq j_v$. Denote by $1 \leq p \leq \infty$ and $1 \leq q \leq \infty$. Let $f : \mathbb{R}^u \to \mathbb{R}$ and $g : \mathbb{R}^v \to \mathbb{R}$ be a couple of real valued bounded functions for which the covariance function is defined. Consider a positive sequence $\eta(r) \to 0$ as $r \to \infty$. The sequences of r.v's $(X_n)_{n\in\mathbb{N}}$, not necessarily stationary, is called conditionally strongly mixing with respect to $\underline{V}_n$ if for all functions f and g the following inequality holds*

$$sup \left| Cov(\mathbb{E}(f(X_{i_1}, \ldots, X_{i_u})|\underline{V}_{i_u}), \mathbb{E}(g(X_{j_1}, \ldots, X_{j_v})|\underline{V}_{j_v}) \right| \leq C(u, v)\eta(r) \left\|f\right\|_p \left\|g\right\|_q, \quad (2.4)$$

here the supremum is taken over all the sequences $(i_1, \ldots, j_v)$ and $\eta(r) \to 0$ as $r \to \infty$ for all $r \geq 1$.

3 Strong mixing conditions for sequence controlled by a factor with long memory

Let $(X_n)_{(n\in\mathbb{N})}$ be a sequence of conditionally independent random variables.
Denote by $\underline{V}_{i_u} = (V_1, \ldots, V_{i_u})$ and $\underline{V}_{j_v} = (V_1, \ldots, V_{j_v})$. Assume that $H1$ and $H2$ are satisfied. We prove the following proposition

Proposition 2 *Let $(X_i)_{i\in N}$ be a sequence of conditionally independent r.v's given $(\underline{V}_i)_{i\in N}$ where $(\underline{V}_i) = (V_1, \ldots, V_i)$. Let $u' \leq u$, $v' \leq v$ and r be positive integers. Let (h_1, h_2) be a couple of real valued bounded functions measurable with respect to $\underline{V}_n$. Assume that the sequence $(V_i)_{i\in N}$ satisfies the following mixing property: for any $u'-$ tuple $(i_1, \ldots, i'_u)$, $v'-$ tuple $(j_1, \ldots, j'_v)$ and any multi-index $i_1 \leq \cdots \leq i_{u'} < i_u + r \leq j_1 \leq \cdots \leq j_{v'}$*

$$sup \left| Cov(h_1(V_{i_1}, \ldots, V_{i_{u'}}), h_2(V_{j_1}, \ldots, V_{j_{v'}})) \right| \leq C'(u, v)\eta(r) \left\|h_1\right\|_p \left\|h_2\right\|_q,$$

where the supremum is taken over all the sequences $(i_1, \ldots, j_v)$ and with $\eta(r) \overset{r\to\infty}{\to} 0$. If in addition the sequence $(X_n)_{(n\in\mathbb{N})}$ is conditionally strongly mixing with respect to $(\underline{V}_n)_{n\in N}$ then for all real valued bounded functions f and g and for any multi-index $i_1 \leq \cdots \leq i_{u'} < i_u + r \leq j_1 \leq \cdots \leq j_{v'}$, the sequence $(X_n)_{(n\in\mathbb{N})}$ is strongly mixing that is

$$sup \left| Cov(f(X_{i_1}, \ldots, X_{i_{u'}}), g(X_{j_1}, \ldots, X_{j_{v'}})) \right| \leq C(u, v)\eta_1(r) \left\|f\right\|_p \left\|g\right\|_q,$$

where the supremum is taken over all the sequences $(i_1, \ldots, j_v)$ and $\eta_1(r) \overset{r\to\infty}{\to} 0$.

3.1 Example

We consider a risk process $(X_i)_{i\in\mathbb{N}}$ defined by $X_i = I_i \times B_i$ where B_i is the claim amount and I_i is a Bernoulli r.v. The B_i's will be considered to be mutually independent and independent of the I_i's and of $\underline{V}$. The structure of dependence of the sequence $(I_i)_{i\in\mathbb{N}}$ is given by the conditional independence property below. Let us denote by $\underline{V}_i = (V_1, \ldots, V_i)$, consider a bounded function h, a constant $K > 0$ and assume:

- for any $i \in \mathbb{N}$, $I_i|\underline{V} \stackrel{\mathcal{L}}{=} I_i|\underline{V}_i$,
- the $I_i|\underline{V}$'s are independent,
- $\mathbb{P}(I_i = 1|\underline{V}_i) = K \sum_{j=1}^{i} \frac{1}{2^{i-j}} h(V_j)$, $\mathbb{P}(I_i = 0|\underline{V}_i) = 1 - \mathbb{P}(I_i = 1|\underline{V}_i)$.

In order to insure $\mathbb{P}(I_i = 1|\underline{V}_i) < 1$ and $\mathbb{P}(I_i = 0|\underline{V}_i) < 1$, we require that $2K \sup h = \kappa < 1$. In this example we take $h(V_j) = (1 + V_j)$, we assume that $(V_n)_{n\in\mathbb{N}}$ is a sequence of independent Bernoulli r.v's and we prove the following proposition

Proposition 3 *If the sequence $(V_i)_{i\in\mathbb{N}}$ is Φ-mixing then, the sequence $(X_i)_{i\in\mathbb{N}}$ is $\Phi(\infty, 1)$-mixing. More generally, it is $\Phi(\infty, v)$-mixing with mixing coefficient*

$$\Phi_1(r) \leq \Phi(\frac{r}{2}) + \frac{1}{\sqrt{2^r}}$$

and $C(v) \leq max[2^v, 2^{v-1}(2\kappa)^v]$.

Remark 2 *If the sequence $(V_i)_{i\in\mathbb{N}}$ is α-mixing then we shall get that $(X_i)_{i\in\mathbb{N}}$ is $\alpha(u, v)$-mixing with mixing coefficient :*

$$\alpha_1(r) \leq 4\alpha(\frac{r}{2}) + \frac{1}{\sqrt{2^r}}$$

and $C(u, v) \leq \max[2^{u+v}, 2^{u+v-1}(2\kappa)^v]$.

3.2 Central limit theorem for sums of conditionally independent random variables given the entire sequence of the factor

In the literature a central limit theorem for α mixing and ρ mixing sequences is proved in Ibragimov (1975) and Ibragimov and Linnik (1971). Similar results were proved earlier by Billingsley (1968) for sequences satisfying Φ mixing condition. Denote by $S_n = \sum_{i=1}^{n} X_i$ and the asymptotic variance of $\frac{S_n}{\sqrt{n}}$ by

$$\sigma^2 = \lim_{n\to\infty} \frac{\mathrm{Var}(S_n)}{n} \tag{3.1}$$

Under some additional conditions, specified in our model, we ensure the existence of σ^2. We provide a self normalized central limit theorem (CLT) for S_n and, following Peligrad and Shao (1995), we introduce the following estimator of σ^2. We denote $S_\ell(j) = \sum_{k=j+1}^{j+\ell} X_k$, $\overline{X}_\ell = \frac{1}{n-\ell+1} \sum_{j=0}^{n-\ell} S_\ell(j)$, $j = 0, \ldots, n - \ell$ and for $\ell_n \to \infty$, $\ell = o(n)$ as $n \to \infty$

$$B_{n,\ell} = \frac{1}{n - \ell + 1} \sum_{j=0}^{n-\ell} \left(\frac{S_\ell(j) - \overline{X}_\ell}{\sqrt{\ell}} \right)^2 ,$$

Theorem 4 *Let* $\{X_n, n \in \mathbb{N}\}$ *be a stationary* $\alpha(4,4)$ *mixing process, satisfying* $\mathbb{E}(X_1) = \mu$ *and* $\mathbb{E}|X_1|^{2+\delta} < \infty$ *for some* $\delta > 0$. *Assume that*

$$\sigma^2 = \lim_{n\to\infty} \frac{Var(S_n)}{n} < \infty \ , \ \sum_{n=1}^{\infty} \alpha(n)^{\delta/(2+\delta)} < \infty \ , \ \ell_n \to \infty \text{ and } \ell = o(n^{\frac{1}{5}}) \text{ as } n \to \infty.$$

Then $B_{n,\ell}$ *is a consistent estimator of* σ^2 *and*

$$\frac{S_n - n\mu}{\sqrt{n}\sqrt{B}_{n,\ell}} \overset{Law}{\to} N(0,1),$$

Remark 3 *In applications, this result can be used for sequences of r.v's converging toward a stationary state. For conditionally independent random variables* $X_{i\in\mathbb{N}}$ *given* $\underline{V}_i$ *this convergence condition is obtained by adding some assumptions on the conditional kernel transition* $N^i(dx_i, \underline{V}_i)$.

References

[1] Cossette, H., Marceau, E. (1999) *The discrete-time risk model with correlated classes of business.* Insurance Mathematics and Economics. 26, 133-149.

[2] Cossette, H., Landriault, D., Marceau, E. (2004) *Compound Binomial risk model in a Markovian environment.* Insurance Mathematics and Economics. 35, 425-443.

[3] Billingsley, P. (1968) *Probability and measure.* Wiley, New York.

[4] Bradley, R.C. (2005) *Basic properties of strong mixing conditions. A survey and some open questions.* Probability Surveys. 2, 107-144.

[5] Dedecker,J. Doukhan,P., Lang, G., León R., JR., Louhichi,S. Prieur, C. (2007) *Weak Dependence: With Examples and Applications.* Lect. Notes Stat. 190.

[6] Dedecker,J. Prieur, C. (2005) *New dependence coefficients. Examples and applications to statistics.* Prob. Th. Rel. Fields. 132 , 203-236.

[7] Dedecker, J. Doukhan, P. (2003) *A new covariance inequality and applications.* Stochastic Process. Appl. 106. 1, 63–80.

[8] Doukhan, P. Louhichi, S. (1999) *A new weak dependence condition and applications to moment inequalities.* Stochastic Process. Appl. 84. 2, 313-342.

[9] Ibragimov, I.A., Linnik, Y.U. (1971) *Independent and stationary Sequences of Random variables.* Wolters-Noordhoff, Groningen.

[10] Ibragimov, I.A. (1975) *A note on the central limit theorem for dependent random variables.* Theory Of Probability and its Applications. Vol. XX, no. 1, 135-141.

[11] Peligrad, M. (1986) *Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables.* Dependence in Probability and Statistics. 193-223.

[12] Peligrad, M. Shao, Q. (1995) *Estimation of the variance of Partial Sums for* $\rho-$ *mixing random variables.* Journal of Multivariate Analysis. no. 52, 140-157.

Théorèmes limites pour des martingales vectorielles à croissance explosive en temps continu et applications statistiques

Hamdi Fathallah * & **Ahmed Kebaier** †

29 mars 2011

Résumé

Dans ce travail, on établit des résultats autour du théorème limite presque-sûre (TLPS) pour des martingales vectorielles quasi-continues à gauche, en temps continu, à croissance *explosive*. Nous exploitons l'ensemble des résultats obtenus dans le cadre d'applications statistiques où nous produisons des vitesses de convergence de type fonctionnel pour l'estimation de paramètre du modèle d'Ornstein-Uhlenbeck bivarié utilisé en modélisation biologique et en mathématiques financières.

Mots Clés : Martingales quasi-continues, théorème de la limite centrale presque-sûre, loi forte quadratique, Ornstein-Uhlenbeck bivarié, normalisation explosive, Estimation de paramètre.

Abstract

In this paper, we establish the almost-sure central limit theorem (ASCLT) for a quasi-left continuous vector martingale with *explosive* growth. We also prove a quadratic extension and establish several new central limit theorems associated with the obtained ASCLT. We finally study the problem of parameter estimation in the particular case of multidimensional diffusion processes, which illustrates in a concrete manner the use of our results.

Keywords : Quasi-left continuous martingales, Almost-sure central limit theorem, Quadratic strong law, Bivariate Ornstein-Uhlenbeck model, Explosive normalization, Asymptotic properties of estimators.

Classification Mathématique. 60G46, 60G51, 60F05.

1 Introduction

On note $\|.\|$ la norme euclidienne sur $\mathbb{R}^d$. Pour une matrice réelle carrée A, A^*, trA et $\det(A)$ désignent respectivement la matrice transposée, la trace et le déterminant de la matrice A. $\mathcal{I}_d$ dénote la matrice identité $d \times d$. La norme de la matrice A est définie par : $\|A\| = \sqrt{tr(AA^*)}$. On note $Vect(A)$, le vecteur obtenu en empilant les vecteurs colonnes de la matrice A et on note $[Vect(A)Vect(A)^*]^{\perp}$ la matrice à blocs dont le bloc d'indice $1 \leq i, j \leq d$ est $A_j A_i^*$ où $A_1, ..., A_d$ sont les vecteurs colonnes de A. Le symbole $\otimes$ désigne le produit tensoriel de mesures ou de matrices.

*Laboratoire LMV, Université de Versailles Saint-Quentin-En-Yvelines, 45 Avenue des Etats-Unis Batiment Fermat 78035 Versailles (France). Tel :+33139253629 ; Fax :+33139254645, E-mail address :hamdi.fathallah@math.uvsq.fr

†Université Paris 13, Institut Galilée, Mathématiques, 99, av. JB Clément, 99430 Villetaneuse, (France), E-mail address : kebaier@math.univ-paris13.fr

Rappelons que toute martingale locale M admet une décomposition unique en $M^c + M^d$, où M^c est la partie martingale locale continue alors que M^d est une somme compensée de sauts nulle en zéro. La variation quadratique de M, notée $[M]$, est le processus défini par

$$[M]_t = \langle M^c \rangle_t + \sum_{0<s\leq t} \Delta M_s \Delta M_s^*,$$

où $\langle M^c \rangle_t$ est l'unique processus croissant continu adapté tel que $MM^* - \langle M^c \rangle_t$ soit une martingale locale nulle en zéro. Le compensateur prévisible du processus $[M]$ est noté par $\langle M \rangle$.

Dans la suite, on considère une martingale quasi-continue à gauche $M = (M_t, t \geq 0)$ d-dimensionnelle, localement de carré intégrable, définie sur un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F})_{t\geq 0}, \mathbb{P})$ (voir Jacod et Shiryaev [2]) et un processus déterministe $V = (V_t)_{t\geq 0}$ à valeurs dans l'ensemble des matrices inversibles. Pour $u \in \mathbb{R}^d$, on définit

$$\phi_t(u) := \exp\left(-\frac{1}{2} u^* \langle M^c \rangle_t u + \int_0^t \int_{\mathbb{R}^d} (\exp\{i\langle u, x\rangle\} - 1 - i\langle u, x\rangle) \nu^M(ds, dx) \right)$$

où ν^M est la mesure de Lévy des sauts de la martingale M. L'ensemble de nos résultats de type presque-sûre est basé sur le théorème de la limite centrale généralisé pour les martingales (voir Touati [4]). Il donne une version généralisée du TLC utilisant non pas la condition *classique* de Lindeberg mais plutôt une hypothèse portant sur les fonctions caractéristiques valable même dans le cas non gaussien.

Théorème 1.0.1 *(Théorème limite central généralisé pour des martingales) Soit $M = (M_t, t \geq 0)$ une martingale locale d-dimensionnelle nulle en zéro et quasi-continue à gauche. Soit $V = (V_t, t \geq 0)$ une famille déterministe de matrices inversibles. Soit $\mathcal{Q}$ une probabilité sur l'espace $\mathrm{C}(\mathcal{X}, \mathbb{R}^d)$ des fonctions continues de $\mathcal{X}$ dans $\mathbb{R}^d$ (où $\mathcal{X}$ désigne un espace vectoriel de dimension finie). Si le couple (M, V) vérifie l'hypothèse suivante*

$$(\mathcal{H}) : \quad \begin{cases} \phi_t((V_t^*)^{-1}u) \longrightarrow \phi_\infty(\eta, u) \;\; p.s., \qquad (t \longrightarrow \infty), \\ \\ \phi_\infty(\eta, u) \;\; \textit{non nulle} \;\; p.s., \end{cases}$$

où η désigne une v.a., éventuellement dégénérée à valeurs dans $\mathcal{X}$. Pour $(z, u) \in \mathcal{X} \times \mathbb{R}^d$,

$$\phi_\infty(z, u) = \int_{\mathbb{R}^d} \exp\{i\langle u, \xi\rangle\} \pi(z, d\xi)$$

désigne la transformée de Fourier des lois conditionnelles unidimensionnelles $(\pi(x, .), x \in \mathcal{X})$ de la probabilité $\mathcal{Q}$. Alors

$$(TLCG) \qquad Z_t := V_t^{-1} M_t \Longrightarrow Z_\infty := \Sigma(\eta) \;\; (t \longrightarrow \infty),$$

de manière stable, où $(\Sigma(z), z \in \mathcal{X})$ est un processus de loi $\mathcal{Q}$ indépendant de la v.a. η.

Sous des hypothèses de régularité pour la normalisation V, nous pouvons obtenir des résultats de type TLCPS à partir du TLCG ci-dessus. Une normalisation V est dite *régulière* si elle vérifie la condition $(\mathcal{C}) = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$ suivante :

- $(\mathcal{C}_1)$ $t \longmapsto V_t$ est de classe $\mathscr{C}^1$.
- $(\mathcal{C}_2)$ il existe $s_0 \geq 0$ tel que pour tout $t \geq s \geq s_0$, on a $V_s V_s^* \leq V_t V_t^*$ (au sens des matrices réelles symétriques positives).
- $(\mathcal{C}_3)$ il existe une fonction $a = (a_t)$ continue sur $\mathbb{R}_+$, décroissante vers 0 à l'infini, telle que

$$A_t = \int_0^t a_s ds \uparrow \infty \quad \text{pour} \quad t \uparrow \infty$$

et une matrice U_1 vérifiant :

$$\begin{cases} a_t^{-1}V_t^{-1}\dfrac{dV_t}{dt} - U_1 = \Delta_{t,1}, & \text{avec } \Delta_{t,1} \longrightarrow 0 \quad (t \longrightarrow \infty), \\ \\ U_1 + U_1^* = S_1, & \text{où } S_1 \text{ est une matrice inversible.} \end{cases}$$

Récemment, Chaâbane et Kebaier [1] ont établi, pour une normalisation V de type régulière et dans le cadre d'obtention du TLCG renforçées par certaines hypothèses, que le couple (M, V) vérifie les résultats ci-dessous. Plus précisément, sous les hypothèses $(\mathcal{H})$,

$$(\mathcal{H}_1) \qquad V_t^{-1}\langle M\rangle_t (V_t^*)^{-1} \longrightarrow C \ \ p.s., \quad (t \longrightarrow \infty),$$

(où C est une matrice aléatoire ou non) et si de plus la condition $(\mathcal{C}_3)$ est obtenue avec

$$\|\Delta_{t,1}\| = \mathcal{O}(t^{-\beta}), \quad (t \to \infty) \quad \text{avec } \beta > 1,$$

ils obtiennent le théorème de la limite centrale presque-sûre généralisé :

$$(TLCPSG) \qquad (\log(detV_t^2))^{-1} \int_0^t \delta_{Z_s}\, d(\log(detV_s^2)) \Longrightarrow \mu_\infty, \ \ p.s., \quad (t \longrightarrow \infty),$$

où μ_∞ est la loi de Z_∞.

Si de plus le couple (M, V) vérifie les hypothèses

$$(\mathcal{H}_2) \qquad V_t^{-1}[M]_t (V_t^*)^{-1} \longrightarrow C \ \ p.s., \quad (t \longrightarrow \infty) \ et$$

$$(\mathcal{H}_3) \qquad C = \int xx^* d\mu_\infty(x),$$

alors ils démontrent une loi forte quadratique :

$$(LFQ) \qquad (\log(detV_t^2))^{-1} \int_0^t V_s^{-1}M_sM_s^*(V_s^*)^{-1} d(\log(detV_s^2)) \longrightarrow C \ \ p.s., \quad (t \longrightarrow \infty).$$

Le but de ce travail est de généraliser ces propriétés à des martingales quasi-continues à gauche à normalisation *explosive*. Nous exploitons l'ensemble des résultats obtenus dans le cadre d'applications statistiques où nous produisons des vitesses de convergence de type fonctionnel pour l'estimation de paramètre du processus d'Ornstein-Uhlenbeck bivarié, utilisé dernièrement pour modéliser le tissu micro vasculaire dans certaines thérapies contre le cancer (voir Favetto et Samson (2008)) et en mathématiques financières (voir les récents travaux de Krämer et Richter (2007) et Lo et Wang (1995)).

2 Enoncés des résultats

Soit $M = (M_t, t \geq 0)$ une martingale locale d-dimensionnelle nulle en zéro et quasi-continue à gauche. Soit $V = (V_t, t \geq 0)$ une famille déterministe de matrices inversibles. Une normalisation V est dite *explosive* si elle vérifie la condition $(\mathcal{C}') = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}'_3\}$ suivante :

- $(\mathcal{C}_1)$ $t \longmapsto V_t$ est de classe $\mathscr{C}^1$.
- $(\mathcal{C}_2)$ il existe $s_0 \geq 0$ tel que pour tout $t \geq s \geq s_0$ on a $V_sV_s^* \leq V_tV_t^*$ (au sens des matrices réelles symétriques positives).
- $(\mathcal{C}'_3)$ il existe une matrice U_2 vérifiant :

$$\begin{cases} V_t^{-1}\dfrac{dV_t}{dt} - U_2 = \Delta_{t,2}, & \text{avec } \Delta_{t,2} \longrightarrow 0 \quad (t \longrightarrow \infty), \\ \\ U_2 + U_2^* = S_2, & \text{où } S_2 \text{ est une matrice inversible.} \end{cases}$$

Ces conditions sont notamment vérifiées dans le cas où V_t est une normalisation scalaire de type $V_t = v_t I_d$ avec v_t une fonction scalaire donnée par $v_t = c\, e^{bt}$ où c et b sont deux réels.

Théorème 2.0.2 *Si le couple (M, V) satisfait aux hypothèses $(\mathcal{H})$ et $(\mathcal{H}_1)$. On suppose que la condition $(\mathcal{C}'_3)$ est obtenu avec*

$$\|\Delta_{t,2}\| = \mathcal{O}(t^{-\beta}), \quad (t \to \infty) \quad \text{avec } \beta > 1,$$

alors on a

$$(TLCPSG) \qquad \mu_t = t^{-1} \int_0^t \delta_{Z_s}\, ds \Longrightarrow \mu_\infty \quad p.s., \quad \text{où } \mu_\infty \text{ est la loi limite de } Z_\infty.$$

Si de plus le couple (M, V) vérifie $(\mathcal{H}_2)$ et $(\mathcal{H}_3)$, alors on a

$$(LFQ) \qquad t^{-1} \int_0^t V_s^{-1} M_s M_s^* (V_s^*)^{-1} ds \longrightarrow C \quad p.s., \quad (t \longrightarrow \infty).$$

Enfin, si le couple (M, V) vérifie

$$(\mathcal{H}''_2) \qquad \exists p \in [1, 2] \ \text{ tel que } \quad \int_0^\infty \int_{\mathbb{R}^d} (1+s)^{-p/2} \|V_s^{-1} x\|^{2p} \nu^M(ds, dx) < \infty \quad p.s.,$$

en plus on suppose que la condition $(\mathcal{C}'_3)$ est obtenue avec

$$\|\Delta_{t,2}\| = \mathcal{O}(t^{-\frac{3}{2}}), \quad (t \longrightarrow \infty),$$

alors on a

$$(TLC) \qquad t^{-1/2} \int_0^t \{U_2 \tilde{D}_s + \tilde{D}_s U_2^*\} ds \Longrightarrow \nu_\infty, \quad (t \longrightarrow \infty),$$

où $\tilde{D}_s = V_s^{-1}(M_s M_s^* - \langle M \rangle_s)(V_s^*)^{-1}$ et conditionnellement à C, ν_∞ est une loi gaussienne matricielle centrée, indépendante de la variable aléatoire C et de covariance

$$\mathscr{C} = (tr(S_2))^{-1} \{2C \otimes C + 2[(Vect(C))(Vect(C))^*]^\perp\}.$$

3 Applications statistiques : modèle d'Ornstein-Uhlenbeck bivarié

Cette application concerne le modèle d'Ornstein-Uhlenbeck bidimensionnel. Ce modèle, plus connu sous le nom de modèle d'Ornstein-Uhlenbeck bivarié, est souvent utilisé notamment en mathématiques financières (voir par exemple Lo et Wang (1995) et Krämer et Richter (2007), mais aussi en biologie où il a permis de modéliser le tissu microvasculaire dans certaines thérapies contre le cancer (voir Favetto et Samson (2008)). Ainsi, les énoncés ci-dessous, viennent compléter les résultats d'estimation des paramètres de ce modèle faits par Favetto et Samson (2008) et Krämer et Richter (2007).

Soit $Z = \{\Omega, \mathcal{F}, (P_z; z \in \mathbb{R}^2)\}$, $F = (\mathcal{F}_t; t \geq 0)$, $(Z_t; t \geq 0)$ une version canonique de la diffusion sur $\mathbb{R}^2$, solution du système différentiel stochastique suivant

$$\begin{cases} dX_t = \theta_1 X_t dt + \theta_2 Y_t dt + dB_t, & X_0 = x, \\ dY_t = \theta_3 Y_t dt + dW_t, & Y_0 = y, \end{cases} \tag{1}$$

où $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ avec $0 < \theta_3 < \theta_1$ et F est la filtration naturelle de $Z_t = (X_t, Y_t)$, P_z est la loi de Z partant de $z = (x, y)$ et $\Gamma = (B, W)$ est un (F, P_z) mouvement brownien plan nul en 0.
En posant $A(\theta) = \begin{pmatrix} \theta_1 & \theta_2 \\ 0 & \theta_3 \end{pmatrix}$, le système (1) s'écrit sous la forme vectorielle

$$dZ_t = A(\theta) Z_t dt + d\Gamma_t; t \geq 0, \quad Z_0 = z. \tag{2}$$

On notera $P_{\theta,z}$ la loi de Z partant de z, pour bien marquer sa dépendance en fonction de θ et $P_{0,z}$ la loi du mouvement brownien Γ. Notons $P^t_{\theta,z}$ et $P^t_{0,z}$ les restrictions de $P_{\theta,z}$ et $P_{0,z}$ à la tribu $\mathcal{F}_t$. D'après le théorème de Cameron Martin Girsanov (voir [3]), on a

$$\frac{dP^t_{\theta,z}}{dP^t_{0,z}} = \exp \mathcal{V}_t(\theta)$$

où

$$\mathcal{V}_t(\theta) = \int_0^t (\theta_1 X_s + \theta_2 Y_s) dX_s + \theta_3 \int_0^t Y_s dY_s - \frac{1}{2}\int_0^t (\theta_1 X_s + \theta_2 Y_s)^2 ds - \frac{\theta_3^2}{2}\int_0^t Y_s^2 ds$$

Soit $D^{(1)}\mathcal{V}_t(\theta)$ la dérivée première de $\mathcal{V}_t(\theta)$ par rapport à θ, on a

$$D^{(1)}\mathcal{V}_t(\theta) = \begin{pmatrix} \displaystyle\int_0^t X_s dX_s - \theta_1 \int_0^t X_s^2 ds - \theta_2 \int_0^t X_s Y_s ds \\ \displaystyle\int_0^t Y_s dX_s - \theta_1 \int_0^t X_s Y_s ds - \theta_2 \int_0^t Y_s^2 ds \\ \displaystyle\int_0^t Y_s dY_s - \theta_3 \int_0^t Y_s^2 ds \end{pmatrix}.$$

L'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ est une solution de l'équation $D^{(1)}\mathcal{V}_t(\theta) = 0$, qui vérifie la relation matricielle suivante

$$M_t = \langle M \rangle_t (\hat{\theta}_t - \theta), \tag{3}$$

où M_t est une $(F, P_{\theta,z})$ martingale définie par

$$M_t = \left(\int_0^t X_s dB_s, \int_0^t Y_s dB_s, \int_0^t Y_s dW_s \right),$$

dont la variation quadratique prévisible est donnée par

$$\langle M \rangle_t = \begin{pmatrix} \displaystyle\int_0^t X_s^2 ds & \displaystyle\int_0^t X_s Y_s ds & 0 \\ \displaystyle\int_0^t X_s Y_s ds & \displaystyle\int_0^t Y_s^2 ds & 0 \\ 0 & 0 & \displaystyle\int_0^t Y_s^2 ds \end{pmatrix}.$$

Dans [4], Touati a déterminé le comportement asymptotique du crochet de M_t à savoir

$$I_t := V_t^{-1} \langle M \rangle_t (V_t^*)^{-1} \longrightarrow I_\infty \quad p.s., \quad (t \longrightarrow \infty),$$

où

$$V_t = \begin{pmatrix} e^{t\theta_1} & 0 & 0 \\ 0 & e^{t\theta_3} & 0 \\ 0 & 0 & e^{t\theta_3} \end{pmatrix}$$

et

$$I_\infty = \begin{pmatrix} \dfrac{1}{2\theta_1} X^2(\theta) & \dfrac{1}{\theta_1 + \theta_3} X(\theta) Y(\theta) & 0 \\ \dfrac{1}{\theta_1 + \theta_3} X(\theta) Y(\theta) & \dfrac{1}{2\theta_1} Y^2(\theta) & 0 \\ 0 & 0 & \dfrac{1}{2\theta_1} Y^2(\theta) \end{pmatrix}.$$

Par conséquent, on a que $I_t V_t (\hat{\theta}_t - \theta) \Longrightarrow \mathcal{N}(0, I_\infty)$. Comme I_∞ est inversible, le TLC précédent s'écrit aussi

(TLC) $$V_t(\hat{\theta}_t - \theta) \Longrightarrow \mathcal{N}\left(0, I_\infty^{-1}\right).$$

Vitesses de convergences pour des fonctionnelles de l'estimateur $\hat{\theta}$ de θ :

La normalisation V_t vérifie la condition $(\mathcal{C}')$ avec $U_2 = \mathrm{Diag}(\theta_1, \theta_3, \theta_3)$ et $\Delta_{2,t} = 0$. Le théorème 2.0.2, appliqué au couple (M, V) définie ci-dessus, permet d'en déduire que l'estimateur $\hat{\theta}$ de θ vérifie les propriétés asymptotiques suivantes

Proposition 3.1 *l'estimateur $\hat{\theta}$ de θ vérifie*

1. $(TLCPS)$ $$t^{-1}\int_0^t \delta_{\{I_s V_s(\hat{\theta}_s - \theta)\}} ds \Longrightarrow \mathcal{N}_{3\times 3}(0, I_\infty) \quad p.s., \quad (t \longrightarrow \infty).$$

2. (LFQ) $$t^{-1}\int_0^t I_s V_s(\widehat{\theta}_s - \theta)(\widehat{\theta}_s - \theta)^* V_s^* I_s^* ds \longrightarrow I_\infty \quad p.s., \quad (t \longrightarrow \infty).$$

   Pour $\tilde{D}_s = I_s V_s(\widehat{\theta}_s - \theta)(\widehat{\theta}_s - \theta)^* V_s^* I_s^* - I_s$, on a

3. $(TLC \ \ de \ \ LFQ)$ $$t^{-1/2}\int_0^t (U_2\tilde{D}_s + \tilde{D}_s U_2)\, ds \Longrightarrow \nu_\infty, \quad (t \longrightarrow \infty),$$

   où conditionnellement à I_∞, ν_∞ est une gaussienne matricielle centrée, indépendante de la v.a. I_∞ et de covariance

$$\mathscr{C} = (2\theta_1 + 4\theta_3)^{-1}\{2I_\infty \otimes I_\infty + 2[(Vect(I_\infty))(Vect(I_\infty))^*]^\perp\}.$$

Références

[1] F. Chaâbane and A. Kebaier. Théorèmes limites avec poids pour les martingales vectorielles à temps continu. *ESAIM Probab. Stat.*, 12 :464–491, 2008.

[2] J. Jacod and A.N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003.

[3] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Foundations, Reprint of the second (1994) edition.

[4] A. Touati. Sur la convergence en loi fonctionnelle de suites de semimartingales vers un mélange de mouvements browniens. *Teor. Veroyatnost. i Primenen.*, 36(4) :744–763, 1991.

Test de comparaison de distributions pour des séquences fortement mélangeantes

Laurence Reboul[1] & Anne-Françoise Yao[2]

1. Universite de Poitiers et Institut de Mathematiques de Luminy, Campus de Luminy, 13288 Marseille Cedex 9, France
reboul@iml.univ-mrs.fr
2. Laboratoire LMGEM, Université Aix-Marseille 2, Campus de Luminy, case 901, 13288 Marseille cedex 09.
anne-francoise.yao@univmed.fr

Résumé. Dans ce travail, nous proposons un test lisse adaptatif de type Neyman pour la comparaison des distributions marginales de deux processus strictement stationnaires et $\alpha-$mélangeants. Nous illustrons les performances de notre test au travers de simulations et applications à des données réelles.

Mots clés : *Tests lisses, processus stationnaire, critère de Schwarz, processus α-mélangeant.*

Abstract : In this paper we propose a Neyman's type smooth test of comparison for the marginal distributions of two strictly stationnary α-mixing sequences. A simulation study and an application to real data show the good performances of our tool.

Key Words : *Smooth test, stationnary process, Schwarz's rule, α-mixing process.*

1 Introduction

La comparaison de deux ou plusieurs séries temporelles est un problème de grand intérêt dans de nombreuses situations pratiques : en écologie par exemple, il est intéressant de comparer l'évolution journalière de constantes atmosphériques telles que l'ozone dans deux régions voisines. En économie, comparer les taux d'intérêt ou les taux d'inflation dans différents pays ou régions du monde peut s'avérer utile pour analyser les politiques économiques et leurs conséquences. Dans la plupart des tests existants, les hypothèses d'indépendance intra ou inter séries sont généralement supposées vraies. Pourtant, ces

propriétés sont rarement réalistes pour la plupart des phénomènes réels. Il suffit de penser par exemple à des séries dont les évolutions sont sous-tendues par un même phénomène, telles que le revenu disponible réel et la consommation réelle, liées toutes deux par la croissance réelle de la productivité ou par les salaires réels. Dans ce travail nous considérons des séries dont la dépendance s'amenuise avec l'éloignement temporel. Plus spécifiquement, nous nous intéressons à des processus $\alpha-$mélangeants.

Considérons deux phénomènes modélisés par des séries temporelles X et Y. Lorsque ces phenomènes sont strictement stationnaires, en notant f_X et f_Y leurs densités respectives inconnues, nous cherchons à tester l'hypothèse non-paramétrique :

$$H_0 : f_X = f_Y \tag{1}$$

contre l'alternative que les deux distributions marginales diffèrent, sur la base d'un vecteur d'observations de chaque processus. Dans de tels cas, on s'attend à obtenir des résultats plus fiables lorsque le test (1) est réalisé via des méthodes spécifiquement conçues pour tenir compte de la dépendance intra-séries.

2 Contexte et statistique de test.

Considérons un processus bivarié strictement stationnaire et α-mélangeant $Z = (Z_t,\, t \in \mathbb{Z})$ défini sur un espace probabilisé $(\Omega,\, \mathcal{A},\, P)$, tel que pour tout $t \in \mathbb{Z}$, $Z_t = (X_t,\, Y_t)$, où (X_t) et (Y_t) sont des processus strictement stationnaires à valeurs réelles de densités respectives f_X et f_Y par rapport à une mesure donnée ν (celle-ci pouvant être choisie de telle sorte qu'elle ne soit pas nécessairement dominée par la mesure de Lebesgue, ce qui permet d'inclure les distributions discrètes). On suppose que ν admet une densité h par rapport à une mesure de reférence λ (qui peut être la mesure de Lebesgue ou une mesure de comptage par exemple). En admettant que f_X et f_Y appartiennent à $\mathbb{L}_2(\nu)$, nous cherchons à tester l'hypothèse non-paramétrique :

$$H_0 \,:\, f_X = f_Y \tag{2}$$

contre l'alternative que les deux distributions sont différentes, à partir d'observations $X^{(n)} = (X_s)_{0\leq s\leq n-1}$ et $Y^{(n)} = (Y_s)_{0\leq s\leq n-1}$ de X et Y respectivement. Ces données peuvent par exemple correspondre aux rendements de n actifs financiers, les indices boursiers étant fixés à plusieurs dates.

Notre test est construit à partir des k $(k > 0)$ plus petits coefficients d'un développement de f_X et f_Y dans une base de polynômes orthogonaux de $\mathbb{L}_2(\nu)$. Plus précisement, notons $(Q_j, j \in \mathbb{N})$ cette base de polynômes. On a alors

$$f_X = \sum_{j\geq 0} a_j Q_j \quad \text{et} \quad f_Y = \sum_{j\geq 0} b_j Q_j,$$

avec $\forall j \in \mathbb{N}$, $a_j = \mathbb{E}(\tilde{Q}_j(X))$, $b_j = \mathbb{E}\left(\tilde{Q}_j(Y)\right)$ et $\tilde{Q}_j = h.Q_j$. Les suites (a_j) et (b_j) tendant vers zéro pour j suffisament grand, on suppose qu'il existe un réel $k > 0$ tel que tester (2) revient à tester l'égalité des vecteurs $(a_j,\ 1 \leq j \leq k)$ et $(b_j,\ 1 \leq j \leq k)$. Une statistique de test peut être naturellement définie de la façon suivante : soit

$$U_n(k) = \frac{1}{\sqrt{n}} \sum_{s=1}^{n} V_s(k)$$

où $V_s(k) = \Big(\tilde{Q}_j(X_s)) - \tilde{Q}_j(Y_s)),\ 1 \leq j \leq k\Big)$. Sous H_0, le processus $U_n(k)$ est de moyenne nulle et de variance $W_n(k) = \mathbb{E}_0\left(U_n(k)U_n(k)'\right)$. Sous certaines conditions portant sur la loi de $V_s(k)$ et sous H_0, $W_n(k)$ converge vers $W(k) = \sum_{t\in\mathbb{Z}} \mathbb{E}_0\left(V_0(k)V_t(k)\right)$. Soit $\widehat{W}_n(k)$ un estimateur consistent de $W(k)$. Si l'inverse de $\widehat{W}_n$ existe alors notre statistique de test s'écrit :

$$T_n(k) = \|\widehat{W}_n(k)^{-1/2} U_n(k)\|^2,$$

La loi asymptotique de $T_n(k)$ sous l'hypothèse d'adéquation est une loi du chi-deux à k degrés de libertés. Ainsi, un test de H_0 consiste à rejeter l'hypothèse nulle dès lors que la statistique dépasse le quantile d'ordre $1-\alpha$ de la loi asymptotique. Le choix de k, important en pratique, est déterminé adaptativement par un critère de type Schwarz.

3 Quelques références

1. Bosq, D., 1989. Test du χ^2 généralisé. Comparaison avec le test du χ^2 classique. Revue de Statistique Applique XXXVII, 43-52.
2. Ghattas, B., Pommeret, D., Reboul, L., Yao, A.F, 2011. Data driven smooth test for paired populations. *Journal of Statistical Planning and Inference.* 141(1), 262-275.
3. Ignaccolo, R., 2004. Goodness-of-t tests for dependent data. Journal of Non- parametric Statistics 16, 19-38.
4. Janic-Wroblewska, J. A., Ledwina, T., 2000. Data driven rank test for two- sample problem. The Scandinavian Journal of Statistics 27, 281-297.
5. Munk, A., Jean-Pierre Stockis, J., Janis Valeinis, J., Giese, G., 2009. Neyman smooth goodness-of-t tests for the marginal distribution of dependent data. Annals of the Institute of Statistical Mathematics, to appear.

Genome

Structural analysis of pocket-ligand pairs, *Stéphanie Perot, Christelle Reynes and Anne-Claude Camproux*

Proteins are associated with fundamental biological processes and their malfunction can engender numerous disease states. One main goal of drug discovery is precisely to prevent, regulate or modify the protein function, by finding a small molecule, a ligand, capable of binding with high-affinity to a specific site on the protein, the binding pocket, and forming a protein-ligand complex. Hence, being able to predict several ligand properties critical for binding to a given pocket, and conversely, is a crucial issue in the field. That is why the present study considers a large set of protein-ligand complexes focusing on binding pocket sites to caracterize different types of pocket-ligand pairs described by both pocket and ligand descriptors. A clustering then provides an original and useful classification which reveals five main different types of pocket-ligand pairs. This classification is then used to propose prediction models of some ligand properties for a given pocket, and conversly which can be very useful in the design of new molecules for drug discovery.

L'analyse d'un réseau de co-expression génique met en valeur des groupes fonctionnels homogènes et des gènes importants relatifs à un phénotype d'intérêt, *Nathalie Villa-Vialaneix, Laurence Liaubet, Thibault Laurent, Adrien Gamot, Pierre Cherel and Magali Sancristobal*

Cet article présente l'analyse d'un réseau de co-expression entre gènes dont la particularité est d'être régulés génétiquement. Cette étude est menée selon deux axes : une classification des gènes impliqués dans le réseau permet de mettre en valeur des groupes fonctionnels homogènes. Par ailleurs, une analyse conjointe du réseau et d'un phénotype d'intérêt permet de mettre en évidence des gènes candidats importants.

A statistic analysis of interactions between serine proteases and inhibitor peptides, *Leslie Regad and Henri Xhaard*

Serine proteases (SP) are proteins involved in biological processes and have a role in diseases such as cancer, inflammatory diseases or coagulation diseases. Therefore, much interest has focused on these proteins for the development of new drugs. In this study, we propose a statistical analysis of interactions between SPs and inhibitor peptides (IP) to characterize the important interactions for complex formation. They allowed the identification of interactions conserved across complexes from different SPs families, and interactions that may occur in specific families. Moreover, these studies allowed the location in the SP active site of important regions for interaction between IPs and SPs.

Use of statistical approach to detect functional motifs in protein loops,

Leslie Regad, Juliette Martin, Gregory Nuel and Anne-Claude Camproux

The determination of protein function is an important challenge in biology. The identification of functional motifs gives useful clues for deducing the protein function. We describe a new protocol to extract motifs of interest from protein loops, based on the structural alphabet HMM-SA and statistical method allowing the extraction of over-represented patterns. HMM-SA allows simplifying three-dimensional structures of loop proteins into sequences of structural letters allowing the application of algorithms dedicated for sequence analysis such as the notion of pattern/word exceptionality. Thus, the statistic over-representation of structural words (short letters sequence) related to SCOP superfamilies is used to extract structural motifs of interest in protein loops. An analysis of the correspondance between over-represented words and biological annotations confirms that some structural motifs strongly over-represented in a SCOP superfamily are involved in the protein function, such as calcium-binding site. Motifs detected by this approach could be used for the annotation of uncharacterized proteins.

Structural analysis of pocket-ligand pairs

Stéphanie Pérot, Christelle Reynès, Olivier Sperandio, Maria Miteva, Bruno Villoutreix & Anne-Claude Camproux

Inserm UMR-S 973, Molécules Thérapeutiques in silico,
Université Paris Diderot, 35 rue Hélène Brion, 75013 Paris, France

Abstract

Proteins are associated with fundamental biological processes and their malfunction can engender numerous disease states. One main goal of drug discovery is precisely to prevent, regulate or modify the protein function, by finding a small molecule — a ligand — capable of binding with high-affinity to a specific site on the protein — the binding pocket — and forming a protein-ligand complex. Hence, being able to predict several ligand properties critical for binding to a given pocket — and conversely — is a crucial issue in the field. That is why the present study considers a large set of protein-ligand complexes focusing on binding pocket sites to caracterize different types of pocket-ligand pairs described by both pocket and ligand descriptors. A clustering then provides an original and useful classification which reveals five main different types of pocket-ligand pairs. This classification is then used to propose prediction models of some ligand properties for a given pocket — and conversly — which can be very useful in the design of new molecules for drug discovery.

keywords: drug discovery, protein-ligand complexes, pocket-ligand pairs

Resumé

Les protéines jouent un rôle fondamental dans les processus biologiques et leur dysfonctionnement est à l'origine de nombreuses maladies. Le but de la découverte de médicaments est justement de prévenir, réguler ou modifier la fonction d'une protéine à l'aide d'une petite molécule appelée ligand, capable de se lier à un emplacement spécifique de la protéine appelé poche, formant ainsi un complexe. Être capable de prédire certaines propriétés d'un ligand capable de se fixer à une poche donnée — et réciproquement — est donc un point crucial. C'est pourquoi nous étudions ici un jeu conséquent de complexes protéine-ligand dans le but de caractériser, à partir de descripteurs de poches et de ligands, les types d'appariements poche-ligand existants. Une classification permet alors de mettre en évidence une décomposition originale en cinq groupes d'appariements poche-ligand. Cette classification permet la mise en place de modèles de prédiction de propriétés de ligands potentiels pour une poche donnée — et réciproquement — informations très utiles pour la conception de nouvelles molécules dans le processus de découverte de médicaments.

1 Introduction

Proteins are associated with fundamental biological processes and their malfunction can engender numerous disease states. One main goal of drug discovery is to found if a small molecule — a ligand — is able to bind to a target protein. In case of structure based drug discovery, one can predict the possible positions and orientations of potential ligands out of a large library of molecules, when it is bound to a specific site on a protein— the binding pocket — to regulate, prevent or modify its function. Although quite efficient this process is long and costly. The reverse process — predicting a potential pocket for a given ligand — is more difficult to address because the binding pocket and even the protein are not known. To address these issues, we present here a multivariate analysis of existing pocket-ligand pairs to define main types of pairs. This analysis results in useful models to predict some ligand properties critical for binding to a given binding pocket and conversely, to predict some properties of a potential binding pocket for a given ligand.

2 Dataset and method

Pairs definition and description

We considere 483 non-redundant three-dimensional high-resolution protein-ligand complexes structures retrieved from two datasets (Wang et al., 2005; Hartshorn et al., 2007). To define pockets, we use the software Surflex (Jain, 2007) — which probes the protein surface with three types of molecular fragments. To describe pocket-ligand pairs, we then compute several pocket and ligand descriptors (Pérot et al., 2010). After redundancy removing, the considered pocket descriptors are volume, roughness, planarity, narrowness, polarity ratio, moments of inertia, hydrogen bond acceptors and donors. The remaining ligand descriptors are volume, moments of inertia, polarity ratio, polar surface area, LogP, rotatable bonds, hydrogen-bonds acceptors and donors.

Pairs visualization and classification

Our analysis is based on the following methods:

- a principal component analysis (PCA) on normalized descriptors which provides a suitable space for our analysis,
- a hierarchical clustering based on a reduced number of the PCA principal components and whose tree allows to determine n pairs clusters,
- analysis of variance — combined with Tukey's Honestly Significant Difference tests — computed on each descriptor against the n clusters to determine whether or not a difference between the clusters exists.

Prediction

We then develop prediction models based on the k-nearest neighbour method applied to the chosen axes of PCA. Two kinds of predictions can be proposed. The first prediction model is done to predict the cluster membership for new pocket-ligand pairs and helps in determining the robustness of our classification. The other prediction models which are more interesting for application purposes are called "pocket-only" and "ligand-only". The "pocket-only" prediction model could be used to predict the properties of a potential ligand for a given pocket. The same process is used with what we call a "virtual ligand" which we create by attributing to each ligand descriptor its mean value in our dataset descriptor i. The "ligand-only" prediction model could be used to predict the properties of a potential pocket for a given ligand. In that case we create what we call a "virtual pocket" in the same way. Concerning the prediction, some pockets and/or ligands — due in particular to flexibility problems — can relevantly be assigned to different clusters. It is then interesting to consider the repartition of pair neighbours with regards to their group membership. Indeed, if a real majority of neighbours belongs to the same cluster, we will decide to affect the unlabelled pair to this cluster — rank 1 prediction. But, if neighbours belong to two clusters with equivalent effectives, the pair will be affected to both clusters — rank 2 prediction — indicating that there might be several possible matchings.

3 Results

Pocket-ligand pairs classification

To propose an accurate description of pocket-ligand pairs we create a clustering tree based on the first fourteen axes of PCA model which captures 95% of the variability. The hierarchical classification tree (Figure 1A) allows to determine five clusters that correspond to five particular pocket-ligand pairs which we also project onto the first two axes of the PCA model (Figure 1B). We show here (Figure 1C and 1D) that a small, rough and not so flat pocket can bind two different profiles of ligands whether the pocket is polar and rather narrow (cluster a) or not (cluster b). Not polar ligands with less hydrogen-bond donors and acceptors correspond with not polar pockets (cluster b). Conversely, a big and not polar pocket can bind two different profiles of ligands depending whether the pocket is rough, flat and narrow (cluster e) or not (cluster d). In the latter the ligand tends to be bigger, more polar, to have more rotatable bonds and hydrogen-bond donors and acceptors (cluster d). The cluster c corresponds to average values for both pocket and ligand properties except for ligand moments of inertia according to the ANOVA results.

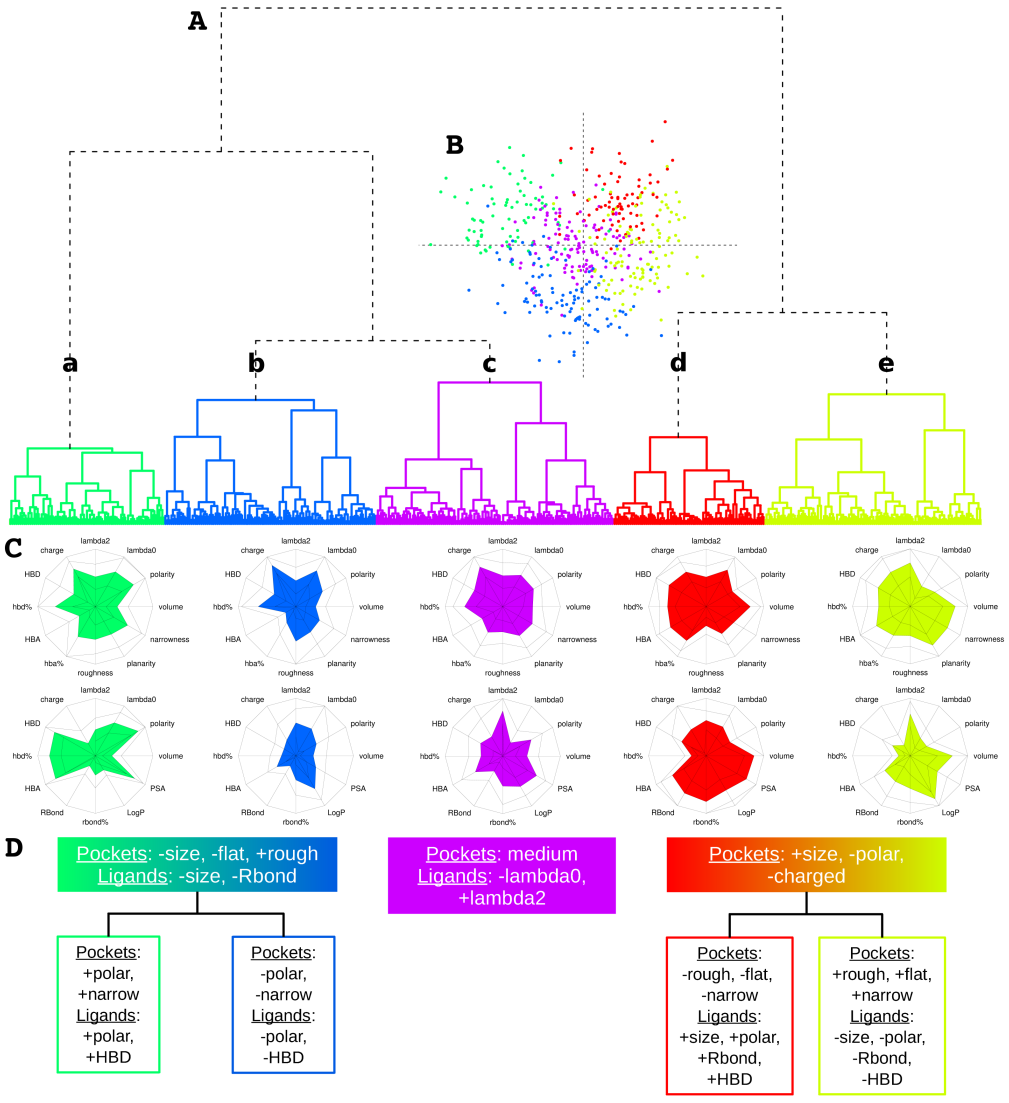


Figure 1: Caracterization of pocket-ligand pairs. A. Hierarchical classification tree. B. Projection onto the two first PCA principal components. C. Star plots of the main pocket (first line) and ligand (second line) properties of each cluster. D. Summary of the main pocket-ligand pairs properties according to the ANOVA results.

Classification relevance

To determine the relevance of our classification we study the repartition of several well-known proteins which can bind several profiles of ligands. Actually, our classification is not only capable of grouping together complexes including a same protein with similar ligands but also to distinguish two complexes of the same protein provided that the ligands are different enough. The cAMP-dependent protein kinase (Figure 2) is given as an illustration: amongst the ten structures of this protein, six out of them are in cluster c (Figure 2A) and four out of them in cluster e (Figure 2B). We can see that the pairs are very similar within a cluster and that they are different between the two clusters although it is the same protein.

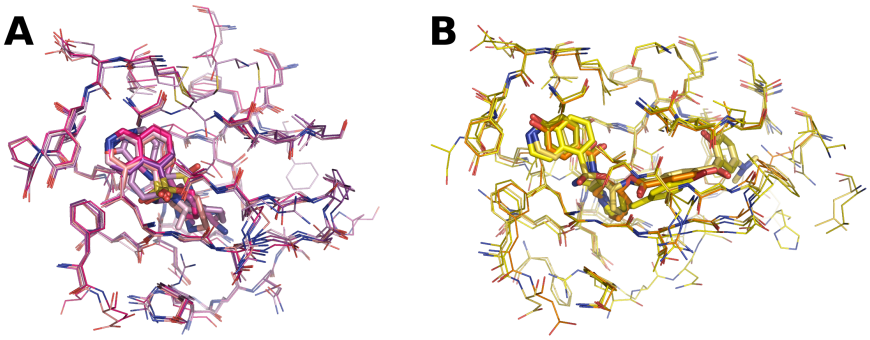


Figure 2: Pocket-ligand representation of the ten crystal structures of cAMP-dependent protein kinase which are in cluster c (A) and cluster e (B). The ligands are depicted in sticks and the pocket amino acids in lines. The pictures are generated with PyMOL.

Prediction

To determine the strength of our prediction models, we first compute a 100-run tenfold cross validation on our dataset (Table 1) with $k = 7$ which appears to provide the best prediction rates overall through robustness studies (results not shown). The proposed model based on pocket-ligand pairs provides good prediction rates which show that our classification is robust. As expected, prediction rates for pocket-only prediction and ligand-only prediction are lower than those for pocket-ligand prediction since less information is taken into account. However these prediction models have more concrete applications for instance in the design of new molecules for drug discovery. The second nearest cluster — rank 2 prediction — is considered when there is a doubt concerning nearest neighbours *i.e.* when there are less than four neighbours in the same nearest cluster. Biologically speaking it is not meaningless since we have shown for instance that a given protein is likely bind two different profiles of ligands and consequently can belong to two clusters.

Prediction method	Rank 1 prediction	Rank 2 prediction	**Total**
Pocket-ligand	97.1%	92.8%	**96.1%**
Pocket-only	73.4%	77.9%	**75.7%**
Ligand-only	85.1%	84.6%	**84.9%**

Table 1: Prediction results based on pocket-ligand prediction, pocket-only prediction and ligand-only prediction using the k-nearest neighbours method with $k = 7$. Rank 1 corresponds to the prediction rates for which there are 5 to 7 neighbours in the same nearest cluster and rank 2 corresponds to the prediction rates on the remaining pairs. Total corresponds to the overall prediction.

4 Conclusion

We present here a new and original way to study binding pockets and reveal five particular types of pocket-ligand pairs. The prediction models give high prediction rates and provide a promising tool which could be very useful for drug discovery. Currently, if one studies a target protein involved in a human disease, this tool could help developping new molecules by predicting its belonging in some of our clusters and so reducing the potential ligand space to look for. Conversely if one knows experimentaly that a ligand is involved in a serie of reactions, this tool can predict some properties of a potential pocket and help in finding which protein can be affected. The next step of this study would be to address the well-known limitation in the field which comes from flexibility of proteins.

References

M. Hartshorn, M. Verdonk, G. Chessari, S. Brewerton, W. Mooij, P Mortenson, and C. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50:726–741, 2007.

A. Jain. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of Computer-Aided Molecular Design*, 21:281–306, 2007.

S. Pérot, O. Sperandio, M. Miteva, AC. Camproux, and B. Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, 15:656–667, 2010.

R. Wang, X. Fang, Y Lu, C. Yang, and S. Wang. The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, 48:4111–4119, 2005.

L'analyse d'un réseau de co-expression génique met en valeur des groupes fonctionnels homogènes et des gènes importants relatifs à un phénotype d'intérêt

Nathalie Villa-Vialaneix[1,2] & Laurence Liaubet[3] & Thibault Laurent[4] & Adrien Gamot[5] & Pierre Cherel[6] & Magali SanCristobal[3]

[1] *Institut de Mathématiques de Toulouse (UMR5219), Université de Toulouse, 118 route de Narbonne, F-31062 Toulouse cedex 9, France*

[2] *Université de Perpignan Via Domitia, IUT, Département STID, Domaine Universitaire d'Auriac, Avenue du Dr Suzanne Noël, F-11000 Carcassonne, France*

[3] *INRA de Toulouse, Laboratoire de Génétique Cellulaire (UMR444), BP 52627, F-31326 Castanet Tolosan cedex, France*

[4] *Toulouse School of Economics, Université Toulouse 1, Manufacture des Tabacs, 21 allées de Birenne, F-31000 Toulouse, France*

[5] *Laboratoire de Biologie Moléculaire Eucaryote (UMR5099), Université Toulouse III, Bâtiment IBCG, 118 route de Narbonne, F-31062 Toulouse cedex 9, France*

[6] *Hendrix Genetics RTC, 100 avenue Denis Papin, F-45808 St Jean en Braye Cedex, France*

Résumé

Cet article présente l'analyse d'un réseau de co-expression entre gènes dont la particularité est d'être régulés génétiquement. Cette étude est menée selon deux axes : une classification des gènes impliqués dans le réseau permet de mettre en valeur des groupes fonctionnels homogènes. Par ailleurs, une analyse conjointe du réseau et d'un phénotype d'intérêt permet de mettre en évidence des gènes candidats importants.

Mots clé : réseau de co-expression génique ; eQTL ; phénotype ; classification ; modularité ; recuit simulé ; diagramme de Moran ; points influents

Abstract

Focusing on genes that are genetically regulated, a gene co-expression network is studied following two purposes: first, the genes are clustered into dense groups that appear to have a great functional homogeneity. Then, jointly studying the network structure and a phenotype of interest, candidate key genes are extracted.

Keywords: co-expression network; eQTL; phenotype; clustering; modularity; simulated annealing; Moran's plot; influential observations

1 Introduction

Nous présentons, dans cet article, une étude de l'expression d'un groupe de gènes contrôlés génétiquement. Cette étude est menée par le biais de la définition puis de l'analyse d'un réseau de co-expression génique et structurée selon deux axes d'intérêt. D'un côté, nous montrons qu'une méthode de classification de sommets dans un graphe permet d'isoler des groupes de gènes homogènes d'un point de vue fonctionnel. L'utilisation d'une telle approche donne des informations aux biologistes sur le rôle de gènes encore inconnus. Par ailleurs, nous expliquons comment l'utilisation d'outils issus de la statistique spatiale permet d'intégrer des informations sur la structure de corrélation des gènes et d'autres relatives à un phénotype d'intérêt. Cela nous conduit à mettre en valeur des gènes potentiellement importants pour ce phénotype.

L'article est structuré de la manière suivante : dans la Section 2, nous présentons les données et le réseau utilisé dans cette étude. La Section 3 présente les résultats de l'analyse du réseau et la Section 4 donne les conclusions de l'étude.

2 Données et définition du réseau de co-expression génique

Les données utilisées dans l'étude présentée ici ont été obtenues à partir de 56 cochons d'une même famille F2. L'expression de gènes a été extraite sur le muscle de la longe des animaux (*longissimus dorsi*) à partir d'une biopuce 9K (GEO numéro d'accession GPL3729). Le protocole d'hybridation et de traitement des données transcriptome (Ferré et al. (2007) ; Lobjois et al. (2008)) a permis l'identification de 2 464 gènes exprimés. A partir des mêmes animaux, l'ADN génomique a été extrait et 170 microsatellites couvrant les 18 autosomes avec un espacement moyen de 17 cM. Les analyses ont été effectuées par SAGA LICOR logiciel. Les animaux F2, leurs parents et grands-parents, ont tous été génotypés et la ségrégation mendélienne a été vérifiée. Les cartes génétiques ont été reconstruites avec le logiciel CRIMAP (Green, 1992). Les matrices de relation IBD (Identity By Descent) ont été estimées tous les 2 cM à l'aide du logiciel LOKI 2,5 (Heath, 1997) et la variance a été estimée à l'aide du maximum de vraisemblance résiduelle (REML) avec la version 2.0 du logiciel ASREML (Gilmour et al., 2006). Ainsi les variations d'expression de 272 gènes ont été identifiées comme étant régulées génétiquement par au moins un locus (ou eQTL, expression Quantitative Trait Locus).

À partir de l'expression des 272 gènes ainsi sélectionnés, un réseau de co-expression génique a été défini en utilisant un modèle graphique Gaussien (Schäfer and Strimmer (2005), parmi les nombreuses méthodes d'inférence de réseau qui existent : voir De Smet and Marchal (2010) pour une revue récente sur la question). 4 000 échantillons bootstrap, chacun de taille 20, ont été utilisés pour estimer les corrélations partielles entre l'expression de toutes les paires de gènes. Le réseau finalement obtenu est modélisé par un graphe de :

- 272 sommets, chacun représentant un gène ;
- 4 690 arêtes entre paires de sommets pour lesquelles la corrélation partielle de l'expression était significativement non nulle (test basé sur une approche bayésienne : Schäfer and Strimmer (2005)) ;
- les arêtes ont été pondérées par la valeur absolue de l'estimation de la corrélation partielle. Les poids des arêtes du graphe sont symétriques et positifs.

3 Analyse du réseau de co-expression génique

3.1 Classification des gènes

Utilisant la structure de corrélation entre expressions, les gènes ont été classés à partir d'un algorithme de classification de sommets (voir Fortunato (2010) pour une revue des méthodes de classification de sommets dans un graphe). L'objectif est d'obtenir des groupes de gènes fortement connectés (c'est-à-dire pour lesquels les corrélations entre expressions sont fortes). De manière plus précise, une mesure de la qualité de la classification de sommets dans un graphe, la modularité, a été optimisée par un algorithme de recuit simulé comme suggéré dans Reichardt and Bornholdt (2006) ou Villa-Vialaneix et al. (2011). La classification obtenue contient 7 classes et conduit à la représentation simplifiée du graphe donnée dans la Figure 1. Cette classification a été confrontée au logiciel

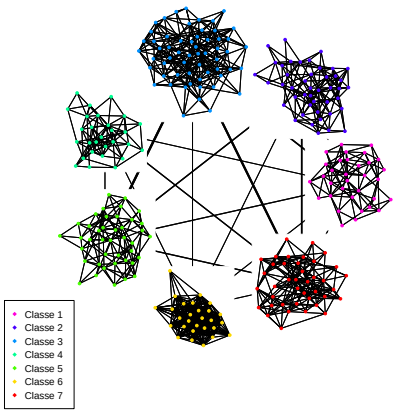


FIGURE 1 – Représentation des classes ainsi que de leurs relations : l'épaisseur des arêtes entre deux classes est proportionnelle à la somme des poids entre les sommets des classes.

"Ingenuity Pathways Analysis" (IPA, https://analysis.ingenuity.com/pa/) pour étudier la pertinence biologique des réseaux de gènes extraits. Les résultats sont donnés dans

Classe	Nombre de gènes dans la classe	Nombre de gènes éligibles pour Ingenuity	Proportion de gènes présents dans un même réseau biologique	Fonctions biologiques correspondantes
1	33	21	94%	Métabolisme des acides gras et des acides nucléiques
2	44	31	96%	Tissus de connexion et morphologie cellulaire
3	58	34	59%	Synthèse protéique et développement musculaire
4	28	27	95%	Prolifération et mort cellulaire
5	41	36	89%	Biochimie et transport moléculaire
6	28	44	63%	Développement et fonction musculaire
7	40	30	75%	Métabolisme des acides aminés et des carbohydrates

TABLE 1 – Résumé de la confrontation des classes trouvées dans le réseau de co-expression génique avec un logiciel d'analyse fonctionnelle.

la Table 1. On remarque une très grande homogénéité fonctionnelle des groupes mis en évidence (le pourcentage de gènes impliqués dans le même réseau fonctionnel est toujours très élevé, sauf pour la classe 3 qui est la plus grande et la moins dense des classes mises en évidence). Ceci plaide en faveur de la pertinence biologique des groupes et peut permettre, par analogie, au biologiste de formuler des hypothèses sur la fonction biologique de gènes inconnus. Dans un contexte où une bonne partie de l'information génétique n'est pas connue, ces hypothèses sont précieuses.

3.2 Lien entre co-expression génique et phénotype d'intérêt

Cette section présente un travail reliant la structure du réseau de co-expression génique à un phénotype d'intérêt impliqué dans la qualité de la viande : le pH. Pour cela, les corrélations partielles entre expression des gènes et valeur du pH de la viande ont été estimées. Les valeurs de ces corrélations partielles sont représentées sur la Figure 2 (gauche) avec des niveaux de couleurs correspondant à la corrélation partielle entre l'expression du gène représenté par le sommet et le pH. L'ajout de cette information supplémentaire peut être modélisée par un réseau dont les sommets sont étiquetées (ici, par la valeur de la corrélation partielle entre l'expression du gène et le pH) : Laurent and Villa-Vialaneix (2010) proposent l'utilisation d'outils issus de la statistique spatiale pour analyser les relations entre la structure d'un graphe et les valeurs des étiquettes sur ses sommets. En particulier, le diagramme de Moran (voir Figure 2, à droite) représente la valeur moyenne des étiquettes des voisins d'un sommet en fonction de la valeur de l'étiquette de ce sommet.

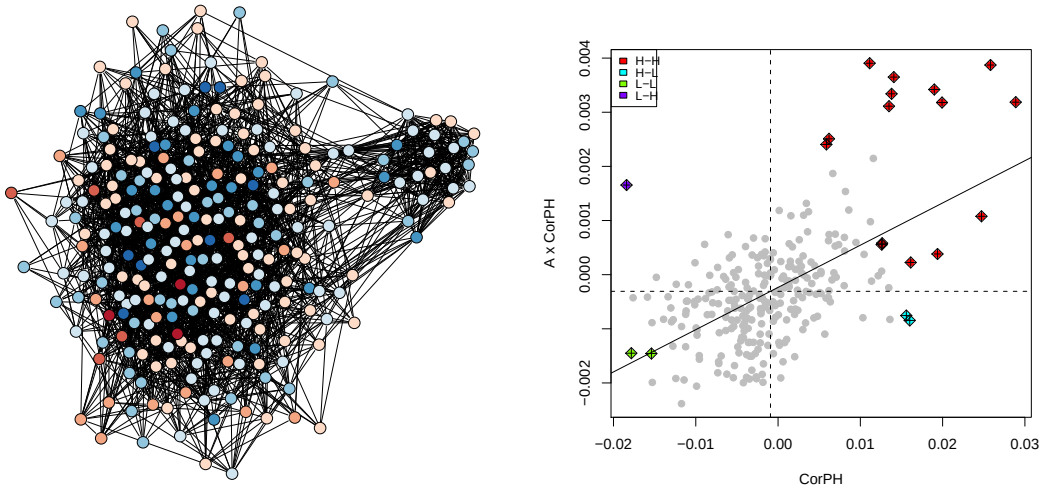


FIGURE 2 – *À gauche* : Niveau des corrélations partielles entre expression des gènes et pH : les sommets bleus sont des gènes dont l'expression est corrélée négativement avec le pH, les sommets rouges des gènes dont l'expression est corrélée positivement. L'intensité de la couleur correspond à la force de la corrélation. *À droite* : Diagramme de Moran des corrélations partielles avec le pH dans le réseau de co-expression génique.

Dans notre cas, ce diagramme présente une tendance linéaire et permet de mettre en valeur des points *influents* : ces points sont ceux qui influencent fortement la valeur de la droite de tendance du diagramme de Moran et dont le comportement peut donc être considéré comme atypique, au sein du réseau, pour la corrélation partielle au pH (voir Belsley et al. (1980) ou Cook and Weisberg (1982) pour plus de détails). La plupart des gènes en rouge, par exemple, correspondent à des gènes dont la corrélation partielle au pH est forte et positive et qui sont entourés de voisins ayant aussi des corrélations partielles avec des valeurs fortes et positives. Beaucoup de ces gènes (10 gènes rouge et le gène violet) se retrouvent dans la classe 4 alors que les autres classes contiennent, au maximum, 5 gènes repérés comme influents. Ainsi, la classe 4 apparaît comme une classe dont la corrélation avec le pH est singulière. Les gènes de cette classe sont des gènes candidats importants pour expliquer le phénotype d'intérêt : ils n'ont pas été sélectionnés par l'approche habituelle consistant à rechercher des différences d'expression selon la valeur du pH car, travaillant sur des gènes dont la particularité est d'être régulés génétiquement, les différences d'expression relatives à un phénotype restent très faibles. De plus, les gènes mis en évidence ne sont pas simplement singuliers par leur relation individuelle au phénotype d'intérêt mais aussi parce que les gènes avec lesquels ils interagissent sont également singuliers.

4 Conclusion

L'utilisation de réseaux est naturelle en biologie puisqu'ils permettent de modéliser des phénomènes de dépendances complexes entre un grand nombre d'objets (ici, des gènes) et donc, de mieux comprendre le système biologique dans son ensemble. Nous avons montré ici que l'utilisation de méthodes statistiques dédiées aux graphes permet de formuler des hypothèses biologiques sur la fonction de certains gènes et de proposer des gènes candidats impliqués dans un phénomène biologique d'intérêt.

Références

Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics*. Wiley, New York.

Cook, R. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8:717–729.

Ferré, P., Liaubet, L., Cocordet, D., SanCristobal, M., Uro-Coste, E., Tosser-Klopp, G., Bonnet, A., Toutain, P., Hatey, F., and Lefebvre, H. (2007). Longitudinal analysis of gene expression in porcine skeletal muscle after post-injection local injury. *Pharmaceutical Research*, 24:1480–1489.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

Gilmour, A., Gogel, B., Cullis, P., and Thompson, R. (2006). *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Green, P. (1992). Construction and comparison of chromosome 21 radiation hybrid and linkage maps using CRI-MAP. *Cytogenetics and Cell Genetics*, 59:122–124.

Heath, S. (1997). Markov chain monte carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, 61(3):748–760.

Laurent, T. and Villa-Vialaneix, N. (2010). Analysis of the influence of a network on the values of its nodes: the use of spatial indexes. In *1ère Conférence Modèles et Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI)*, Toulouse, France.

Lobjois, V., Liaubet, L., SanCristobal, M., Glenisson, J., Feve, K., Rallieres, J., Le Roy, P., Milan, D., Cherel, P., and Hatey, F. (2008). A muscle transcriptome analysis identifies positional candidate genes for a complex trait in pig. *Animal Genetics*, 39(2):147–162.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(016110).

Schäfer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Villa-Vialaneix, N., Dkaki, T., Gadat, S., Inglebert, J., and Truong, Q. (2011). Recherche et représentation de communautés dans un grand graphe : une approche combinée. *Document Numérique*, 14(1):59–80.

A statistic analysis of interactions between serine proteases and inhibitor peptides

Leslie Regad[1,2] & Henri Xhaard[3]

1: INSERM, U793, Paris F-75013, France
2: Université Paris 7 - Paris Diderot, UMR-S793, MTi, F-75013 Paris, France
3: CDR, University of Helsinki, P.O. Box 56, FI-00014 Helsinki

Abstract

Serine proteases (SP) are proteins involved in biological processes and have a role in diseases such as cancer, inflammatory diseases or coagulation diseases. Therefore, much interest has focused on these proteins for the development of new drugs. In this study, we propose a statistical analysis of interactions between SPs and inhibitor peptides (IP) to characterize the important interactions for complex formation. They allowed the identification of interactions conserved across complexes from different SPs families, and interactions that may occur in specific families. Moreover, these studies allowed the location in the SP active site of important regions for interaction between IPs and SPs.

Resumé:

Les protéases à sérine (SP) jouent un rôle dans les maladies tels que les cancers, les maladies inflammatoires ou de coagulation. C'est pourquoi, un grand intérêt s'est porté sur ces protéines pour le développement de nouveaux médicaments. Dans cette étude, nous proposons une analyse statistique des intéractions entre les SPs et leur peptides inhibiteurs (IP) afin de caractériser les interactions importantes pour la formation des complexes. Elles ont permis d'identifier les interactions conservées au seins des complexes provenant de différentes familles de SPs, et des intéractions pouvant intervenir dans la spécificité des familles. De plus, ces analyses ont permis de localiser dans le site actif des SP, les régions importantes pour l'interaction entre les IP et les SPs.

keywords: Serine proteases, peptide inhibitors, data mining, multivariate analysis

1 Introduction

Serine Protease (SP) belongs to the family of proteases (enzymes that cut protein bonds). They are presents in unicellular and complex organisms and often involved in multiple biological processes: digestive, blood clotting, immune system, and inflammation processes Rawlings and Barrett (2000); Coughlin (2000). Therefore,

SPs have often actions in cancer, inflammatory and cardiovascular diseases Rawlings and Barrett (2000); Coughlin (2000). Thus, many SPs are a major focus of attention for pharmaceutical industry as potentional drug targets.
Some inhibitor peptides (IP) of SPs have been identified Di Marco and Priestle (1997); Luckett et al. (1999). The use of these peptides for the developement of new drugs is an important challenge. However, there are several disadvantages associated with the use of peptides as drugs: their limited stability towards proteolysis, their poor transport properties, their rapid excretion, and their inherent flexibility that enables interaction with other targets. One way to avoid these disadvantages is to design a peptidomimetic compound. A peptidomimetic is a compound containing non-peptidic structural elements that is capable of mimicking the biological actions of a natural parent peptide and that does no longer have classical peptide characteristics such as enzymatically scissile peptidic bonds.
In this study, we present statistical analyses of interactions between SPs and inhibitor peptides in order to the identify the important interactions for the binding between SPs and IPs. This knowledge will facilitate the construction of stable and active peptidomimetic compund inhibiting SPs.

2 Material & Methods

2.1 Data

A set of 71 complex SPs / IPs has been extracted from the Protein Data Bank (PDB).

2.2 Extraction of atoms involved in interactions proteins/peptides

Protein atoms involved in interaction with peptide were defined as protein atoms located at less than 3.5 Å far from peptide atoms. The set of these selected atoms consists to pocket atoms.

Peptide atoms involved in interaction with protein were defined as peptide atoms located at less than 3.5 Å far from pocket atoms.

For each complex, we extracted all interactions. The set of extracted interactions was also classified according their type (protein atom-peptide atom). For each interaction type we computed the number of times it occurs in the dataset, named n_{obs}.

2.3 Specificity of interaction of each family

From the matrix containing the number of each interaction type for all SP families, we performed a chi-squared test and a correspondence analysis in order to extract the interactions common or specific to different SP families.

2.4 Analysis of significance of interaction

In order to know if an interaction is random or not, we compared its occurrence (n_{obs}) in our data set to its occurrence in a random dataset. To perform that, we used a procedure similar to ones presented in Martin (2010). The random dataset is built by permuting the pocket atom set and the ligand atom set (involved in interactions). From this random set of interactions, we computed the occurrence of each interaction, named N_{rand}. The random procedure was repeated 1000 times in order to obtain the distribution of each interaction occurrence under the null hypothesis of random counts. The significance of each interaction could be assessed by computing the corresponding p-value defined by :

$$p - value = P(N_{rand} \geq n_{obs}) \tag{1}$$

A high p-value indicates that the interaction has a high probability to occur by chance. Conversely, a very low p-value means that interaction is significantly higher than expected with a random model, and results of specificities of proteins and peptides. A value of 5% is classically used as significant cut-off and we used the FDR correction to take into account the multiple tests.

2.5 Density map of atoms involved in interactions

For pocket and ligand atoms we represented the density map. For this, the 3D coordinates of all atoms are extracted from the pdb files of each protein. Then, we estimated the density of atoms using a method based on the kernels. To perform that, we used the package *feature* of the R software allowing the multivariate kernel density estimation (Duong et al., 2008). Then, we represented the 3D density map of atoms using the *plot* function available in the *feature* package.

3 Results & Discussion

We disposed of a set of 71 complexes of SPs bound to a peptide in their active site. Complexes cover 11 SP families (acrosin, cathepsin, chymase, chymotrypsin, elastase, enteropeptidase, factor x or ix or vii or xi, granzyme, hepsin, kallikrein, matriptase, plasminogen-activator, prostatin, thrombin, trypsin, tryptase, urokinase) and different species: *Homo sapiens, Bos taurus, Sus scrofa*, ... The peptide lengths range from 3 amino acids to 50 for the largest one.

From the set of SP/IP complexes, we extracted 2,535 interactions. Each interaction is named by "protein atom – peptide atom", for example interaction $O_{ms} - N$ corresponds to an interaction between a oxygen atom of the main chain of protein with an azote atom of peptide. In the interaction set, we counted 28 different types of interactions. On average, a SP has 36 interactions with a peptide ($\pm$ 14), with a minimum one interaction, and a maximum of 79 interactions. It is clear that the number of interactions by complex strongly depends on the length of peptide: the longer the peptide, the larger extracted-interaction number.

3.1 Specificity of interactions among each SP family

We computed the occurrence of interaction types in each SP family. In order to know if the repartition of interactions is different across SP families we performed a chi-square test. We obtained a significative test (p-value=$6.722e^{-06}$), showing that the repartition of interactions differs according to SP families. This result must be taken with caution because lot of interactions have an expected occurrence below 5. To vizualise the specificities of interactions of each SP family, we performed a correspondence analysis (CA). The first two components represent 56% of the variability. Figure 2 presents the projection of the SP families and the interaction types on those first two components. At first, we can observe that in terms of interactions, we can group (i) urokinase family and thrombin family, (ii) trypsin, kallikrein, trypstase, factor-ix. Elastase, granzyme and enteropeptidase seem to be particular in terms of interactions. Thanks to CA plot, we identify the interaction types likely to be common to the different SP families (interaction types located in the plot center). Thus we can conclude that interactions $N - O, C_{mc} - O, O_{mc} - O, O_{mc} - C, O - N$ and $O - C$ are present in all SP families, in agreement with the large occurrence of these interactions in SP families. In parallel, we identify the interaction types specific to one or several SP families. For example, the interaction between a sulfur protein atom and an oxygen ligand atom $S - O$ is seen only on granzyme and enteropeptidase families. The interaction between a carbon protein atom and a carbon ligand atom $C - C$ is specific to granzyme and elastase families. Therefore, we can suppose that these interactions are involved in the specificites of each family.

3.2 Significativity of interactions

The second question that we asked is: Do the extracted interactions result from the complex formation or do they occur by chance ?. To answer this question, we analysed the significance of each interaction by comparing their occurrence in the dataset to ones computed in a random dataset (see Method). We counted nine significant interaction types ($C\alpha - O, C - O, C_{mc} - C, N - O, N_{mc} - S, O - C, O - N, O_{mc} - C, O_{mc} - N$) in the set of 71 peptides. We can conclude, that these interaction types are very important for the interaction between proteins and peptides.

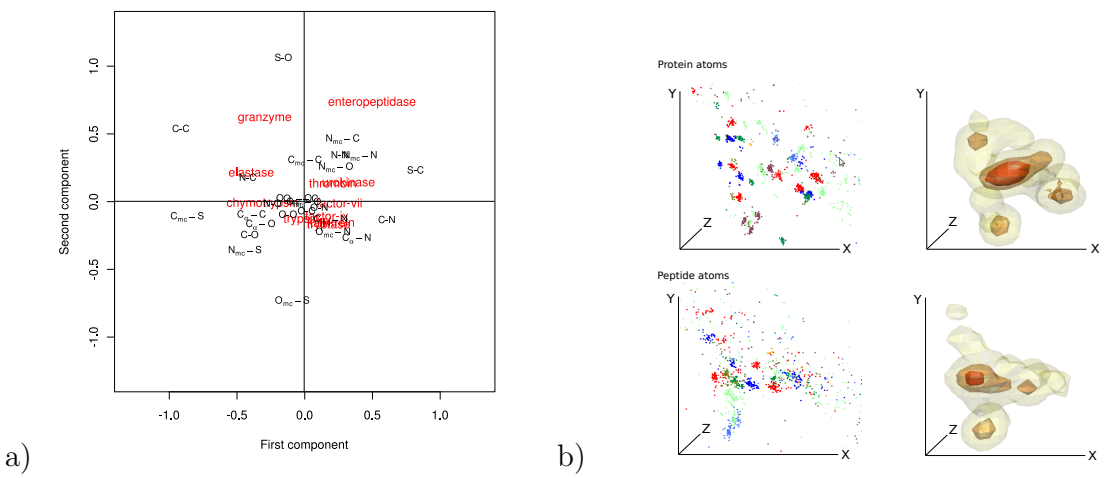


Figure 1: Analysis of interactions extracted from SP/IP complexes. a) corresponds to the correspondence analysis of interaction counts in SP families. In red are colored the SP families, in black are colored interaction types. b) Visualisation of the conservation of atoms involved in interaction between SPs and IPS. Top panel corresponds to pocket atoms. Bottom panel corresponds to peptide atoms. Left panel corresponds to density map of atoms. Right panel corresponds to 3D coordinates of atoms.

Moreover, from this set of significant interaction types, four correspond to interactions previously identified as common to the different families.
In a second step, we extracted significant interactions from the four SP families (trypsin, thrombin, elastase, chymotrypsin) with more than 2 complexes. We can observe, that interactions $N - O$ and $O - N$ are significant in the four SP families, showing that these interactions are very important for the interaction between proteins and peptides.

3.3 Location of conserved interactions

We represented pocket and peptide atoms involved in interactions on the 3D plan according to their X, Y, and Z-coordinates. Two types of graphs were performed: (i) representation of atom coordinates in the 3D space for atoms extracted from all families, and (ii) the 3D density map of atoms computed from their 3D coordinates. Figure 1b presents the location of protein atoms involved in interactions with peptides. The density maps shows that four populated regions can be distinguished, particularly one very dense region in the center of the plot. The visualization of the atom types shows that in conserved regions, atoms are grouped according their type.

We performed the same analysis with the peptide atoms involved in the interactions with SP, see Figure 1b. On the density map, three conserved regions are distinghuished, where atoms are grouped according to their types. However, contrary to pocket atoms, the atom clusters are less well-defined, showing that pocket atoms are more conserved than peptide atoms
Thus, these graphics allows us to identify important regions for the interactions between SPs and IPs.

4 Conclusion

We present a statistical analysis of interactions between SPs and their IPs. They allowed the identification of interaction types common to SP families and putative specific interactions. Moreover, we showed that interactions between azote atoms and oxygen atoms are very important for the contact between SPs and IPs in different families. The vizualisation of atoms involved in interactions in a tri-dimensional space allows the identification of conserved interactions across SP families. These knowledges could be use to create new peptidomimetic compounds inhibiting SPs with a better stability and more effective, and more specific to a given SP family.

References

SR. Coughlin. Thrombin signalling and protease-activated receptors. *Nature*, 407: 258–64, 2000.

S. Di Marco and J.P. Priestle. Structure of the complex of leech-derived tryptase inhibitor (ldti) with trypsin and modeling of the ldti-tryptase system. *Structure*, 5:1465–1474, 1997.

T. Duong, A. Cowling, I. Koch, and M.P Wand. Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 52:4225–4242, 2008.

S. Luckett, R.S. Garcia, J.J. Barker, A.V. Konarev, P.R. Shewry, A.R. Clarke, and R.L. Brady. High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *J.Mol.Biol*, 209:525–533, 1999.

J. Martin. Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. *PLoS Comput Biol*, 6:e1000821, 2010.

N.D. Rawlings and A.J. Barrett. Merops: the peptidase database. *Nucl Acids Res*, 28:323–325, 2000.

Using statistical approach to detect new functional structural motifs in protein loops

Leslie Regad[1,2] & Juliette Martin[3] & Grégory Nuel[4] & Anne-Claude Camproux[1,2]

1: INSERM, U793, Paris F-75013, France
2: Université Paris 7 - Paris Diderot, UMR-S793, MTi, F-75013 Paris, France
3: Université Lyon 1, Univ Lyon, France; CNRS, UMR 5086 ; Bases Moléculaires et Structurales des Systèmes Infectieux, IBCP 7 passage du vercors, F-69367, France
4: MAP5, UMR CNRS8145, Universit Paris-Descartes, Paris

Abstract

The determination of protein function is an important challenge in biology. The identification of functional motifs gives useful clues for deducing the protein function. We describe a new protocol to extract motifs of interest from protein loops, based on the structural alphabet HMM-SA and statistical method allowing the extraction of over-represented patterns. HMM-SA allows simplifying three-dimensional structures of loop proteins into sequences of structural letters allowing the application of algorithms dedicated for sequence analysis such as the notion of pattern/word exceptionality. Thus, the statistic over-representation of structural words (short letters sequence) related to SCOP superfamilies is used to extract structural motifs of interest in protein loops. An analysis of the correspondance between over-represented words and biological annotations confirms that some structural motifs strongly over-represented in a SCOP superfamily are involved in the protein function, such as calcium-binding site. Motifs detected by this approach could be used for the annotation of uncharacterized proteins.

Resumé:

La détermination de la fonction des protéines est un enjeu actuel important en biologie. L'identification de motifs fonctionnels est un moyen efficace d'inférer la fonction des protéines. Nous décrivons ici un protocole, basé sur l'alphabet structural HMM-SA et une approche statistique de recherche de pattern sur-représentés, pour extraire des motifs structuraux fonctionels dans les boucles protéiques.

HMM-SA permet de simplifier les structures tri-dimensionnelles des boucles en des séquences uni-dimensionnelles de lettres, permettant l'utilisation des méthodes dédiées à l'analyse des séquences protéiques pour l'extraction de motifs structuraux. Ainsi, la sur-représentation statistique des mots structuraux (séquences de lettres structurales) dans les superfamilles de la classification de SCOP est utilisée pour extraire des motifs structuraux statistiquement exceptionnels dans des superfamilles de protéines associées à des fonctions particulières.

Une analyse de la correspondance des mots les plus sur-representés avec des annotations biologiques confirme que ces motifs sont impliqués dans la fonction des

protéines, comme par exemple dans les sites de fixation au calcium. Ces résultats illustrent que les motifs détectés par cette approche pourront être utilisés pour l'annotation des protéines non caractérisées.

keywords: structural alphabet, functional motifs, protein loops, statistic over-representation, data mining

1 Introduction

The prediction of protein function is a very important challenge. The identification and prediction of functional sites can give useful clues for deducing the protein function. Two types of methods have been developed for binding site prediction: some exploit the conservation of amino-acid sequence of binding sites (Andreini et al., 2004; Shu et al., 2008), others exploit the three-dimensional (3D) structure conservation of binding sites (Polacco and Babbitt, 2006; Bordner, 2008). Most of these methods need for the learning functional motifs the knowledge of the position of functional site, and the computation of structural alignment or geometric descriptors.

In this paper, we present an alternative strategy for functional motif identification, based on the structural alphabet HMM-SA (Camproux et al., 2004) and on statistic over-representation (Nuel et al., 2010). HMM-SA is a collection of 27 structural prototypes of four residues, called structural letters, established after a structural classification of four-residue fragments. It allows the simplification of all protein structures into uni-dimensional structural-letter sequences. In a previous study, we have shown structural words, successive four structural letters, extracted from structural-letter sequences of loop structures correspond to clusters of seven-residue fragments with similar structures and amino-acid preferences (Regad et al., 2010). We investigate the link between structural words and protein function by looking for words over-represented in superfamilies as defined by SCOP classification (Murzin et al., 1995). The role of these motifs in the protein function is then analyzed using the biological annotation Swiss-Prot.

2 Extraction of structural words with a putative functional role

We assumed that a structural motif with an important role in protein function has been conserved during evolution resulting in its over-representation in the set of proteins associated to this functional role. We used the structural classification of proteins SCOP that groups at superfamily level proteins with similar structures and functions. Thus, structural motifs over-represented in a SCOP superfamily are potential structural motifs of interest, which could be involved in protein function.

Extraction of structural words

The protocol to extract structural motifs is detailled in Regad et al. (2010). It is based on the structural alphabet HMM-SA. It is a set of 27 prototypes of four residues, established by hidden Markov models allowing the classification of four-residue fragments extracted from proteins according their geometry (Camproux et al., 2004). Thanks to HMM-SA, a structure of n residues is simplified into a sequence of $(n-3)$ structural letters, where each structural letter describes the conformation of four-residue fragments. The simplification is achieved by a dynamic programming algorithm based on Markovian process to obtain maximum *a posteriori* encoding using the Viterbi algorithm and takes into account the geometry similarity between four-residue fragments and the 27 structural letters.

A set of 8 119 non-redundant proteins were simplified into structural-letter sequences using HMM-SA. The structural-letter sequences corresponding to loop structures were further split into overlapping words of four letters corresponding to structural motifs of seven-residue fragments (Regad et al., 2010). From the 90 811 loops extracted from the protein data set, we extracted 25 304 different structural words describing the conformation of 238 158 seven-residue fragments.

Computation of structural word over-representation

We computed the over-representation of structural words in each SCOP superfamily corresponding of structural-letter sequences of loops belonging to the superfamily. As structural-letter sequences of loops correspond to short sequences, the classical methods for the word over-representation computation did not applicable. We used the SPatt software (Nuel et al., 2010) to compute the exact statistics of structural words, because SPatt allows the computation of exact statistic of words extracted from a data set composed of a large number of short sequences. The over-representation computation is achieved by comparing the real frequency (n_w) of a given word w in a superfamily and its expected ones (N_w) computed under a background model defined as a first order Markov chain estimated on the set of protein loops. The over-representation score of a word w is given by:

$$Lp(w) = -\log_{10}[P(N_w > n_w)]$$

where P the probability of the event. A Lp score equal to 3 means that the pattern is over-represented with a p-value of 10^{-3}. To define the type of a word is over-represented, we compared its Lp to a threshold. This significance threshold was defined by taking into account the multiple testing and was set to 5.97 using Bonferroni correction.

3 Results

3.1 Words strongly over-represented in a superfamily seems to be involved in protein function

Among the 11 494 structural words, we counted 1 705 over-represented words in at least one superfamily. From this over-represented word set, 23 are strongly over-represented in at least five superfamilies, with a Lp higher than 50. Thus, these words are specific to the superfamily, and could be involved in protein function. To confirm this hypothesis, we analyzed the correspondence between these 23 structural words and biological annotation extracted from Swiss-Prot database. It is a curated sequence database which strives to provide a high level of functional annotations (Bairoch et al., 2005) for proteins. This comparison is under-estimated because Swiss-Prot database is not complete. For example, only 2% of the set of seven-residue fragments are annotated by biological annotations. This analysis allows to confirm that five structural words contain residues involved in binding site. We illustrated this analysis using the word `DODQ` that contains residues involved in calcium-binding site.

3.2 The `DODQ` word corresponds to calcium-binding site

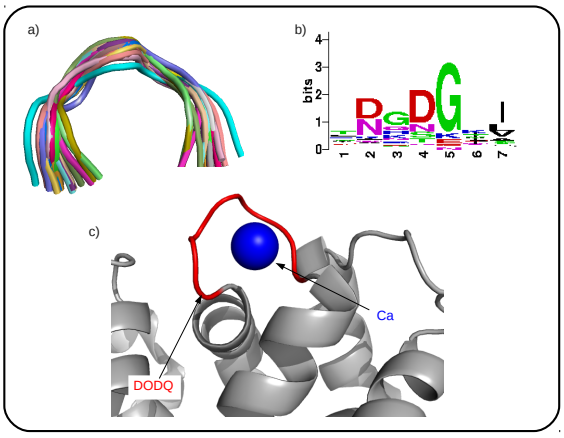


Figure 1: Illustration of the functional word `DODQ` with a) the superimposition of associated structural fragments, b) the amino-acid conservation using a Logo (Crooks et al., 2004), and c) an example of the word in a protein structure. Motifs and protein structure are represented using Pymol (DeLano).

The structural word `DODQ` is seen 73 times in loop dataset. Figure 1a presents the geometry of 30 superimposed `DODQ`-fragments and shows that they have similar

geometry. This word presents strong amino-acid specificities as presented in Figure 1b. It is strongly over-represented (over-representation score = 157) in the superfamily "EF-Hand" (SCOP id = 47473) grouping proteins with calcium-binding site. Thus, we can suppose that DODQ word is involved in calcium-binding site of proteins of the superfamily "EF-Hand".

To confirm this hypothesis, we analyzed the correspondence between DODQ and the Swiss-Prot annotation "calcium-binding site", allowing the localisation of the "calcium-binding site" in proteins. In our dataset, we counted 20 "calcium-binding site" annotations extracted from 16 proteins. We extracted 23 DODQ-fragments from these 16 proteins, which 15 are annotated by "calcium-binding site". This confirms that structural word DODQ is very strongly associated to the calcium-binding site annotation with a precision of 65%. Moreover, these results suggest that nine unannotated fragments are calcium-binding sites.

To investigate the functional role of these unannotated fragments, we predicted the calcium-binding sites from proteins containing these unannotated fragments using the software SitePredict (Bordner, 2008). This is a machine learning-based method based on diverse residue-based properties including spatial clustering of residue types and evolutionary conservation. Six out of the nine unannotated DODQ-fragments contain residues predicted as involved in calcium-binding sites. These results allows to conclude that structural word DODQ are involved in calcium-binding sites.

Lastly, we analyzed the capacity of DODQ for detecting calcium-binding sites. To perform that, we computed the sensibility of DODQ to detect calcium-binding sites, that means the proportion of calcium-binding sites detected by the word DODQ. We counted that 75% of calcium-binding sites (15/20) extracted from proteins are detected by DODQ. These results show that structural word DODQ allow a good detection of calcium-binding site.

4 Conclusion & Perspectives

We present a method allowing the extraction of 3D motifs from loops involved in protein function. This method is based on the simplification of loop structures into strings using the structural alphabet HMM-SA and the over-representation of motifs in a set of proteins with similar function, provided by SCOP superfamily classification.

The analysis of the functional annotations, provided by Swiss-Prot database, of motifs strongly over-represented in a superfamily showed that some of these motifs are involved in binding sites of small ligands. This confirms that statistical over-representation of HMM-SA simplified motifs is an effective tool for the extraction of motifs of interest involved in protein function. Thus, functional words could be used to improve the functional annotation of proteins and quickly complete annotations in Swiss-Prot database. Moreover, the identification of these functional words in

structural-letter sequences corresponding to the structure of uncharacterized proteins is useful for the prediction of functional sites in these proteins.

References

C. Andreini, I. Bertini, and A. Rosato. A hint to search for metalloproteins in gene banks. *Bioinformatics*, 20(9):1373–1380, 2004.

A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. The universal protein resource (UniProt). *Nucl Acids Res*, 33:154–159, 2005.

A.J. Bordner. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24):2865–2871, 2008.

A. C. Camproux, R. Gautier, and T. Tufféry. A hidden Markov model derivated structural alphabet for proteins. *J Mol Biol*, 339:561–605, 2004.

G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: A sequence logo generator. *Genome Res*, 14:1188–1190, 2004.

W. L. DeLano. The pymol molecular graphics system (2002) on world wide web http://www.pymol.org.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, 1995.

G. Nuel, L. Regad, J. Martin, and A. C. Camproux. Exact distribution of pattern in a set of random sequences generated by a Markov source: application to biological data. *Algo Mol Biol*, 5:15, 2010.

B. J. Polacco and P. C. Babbitt. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, 22:723–730, 2006.

L. Regad, J. Martin, G. Nuel, and A. C. Camproux. Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, 11:75, 2010.

N. Shu, T. Zhou, and S. Hovmoller. Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, 24(6):775–782, 2008.

Index