Mercredi 25 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 communications 189

Résumés des conférenciers invités

Mercredi 25 Mai 2011, 8h30-9h15

SECOND ORDER BSDES AND MONTE CARLO METHODS FOR FULLY NONLINEAR PDES

Nizar Touzi, CMAP, Ecole Polytechnique Paris,

Backward stochastic differential equations (BSDEs) appeared in Bismut [1] in the linear case, and received considerable attention since the seminal paper of Pardoux and Peng [7]. The various developments are motivated by applications in probabilistic numerical methods for partial differential equations (PDEs), stochastic control, stochastic differential games, theoretical economics and financial mathematics.

CRITICAL ISSUES AND DEVELOPMENTS IN HIGH-DIMENSIONAL CLASSIFICATION IN BIOMEDICAL RESEARCH

Anne-Laure, Boulesteix Université de Munich

In the first part of this talk, I give an overview of state-of-the-art methods for classification with high-dimensional biological data in small sample settings. I also discuss current challenges and perspectives for statistical research including the design of very sparse prediction rules, the validation of the obtained rules and of their added predictive value (Boulesteix & Sauerbrei, 2010), the stability of prediction rules and gene lists as reviewed in Boulesteix & Slawski (2009), and global tests for high-dimensional data as suggested in Boulesteix & Hothorn (2010).

The second part of the talk is devoted to the correction of optimization biases in the context of error estimation through cross-validation (CV). Most modern classification and regression methods for small sample high-dimensional data incorporate a tuning parameter that adjusts the complexity to the considered data set. The choice of this parameter should be given much attention. Simply reporting the performance of the "best" tuning parameter leads to over-optimistic conclusions in the sense that the reported prediction error is downwardly biased. A similar problem occurs when researchers apply different prediction methods successively and report the smallest prediction error only. In this case, the tuning parameter is the method itself rather than a real-valued parameter, but the bias is of the same nature. These problems affect small sample high-dimensional data particularly due to 1) the high variance of the error estimates, 2) the multiplicity of available methods, 3) the relative lack of standards and diagnostic tools.

Posters :

Posters : MERCREDI 25 MAI 2011, 9h15-10h50

Announcement effect and intraday volatility patterns of euro-dollar exchange rate : monetary policy news arrivals and short-run dynamic response, *Mokhtar Darmoul and Mokhtar Kouki*

Dans cet article, nous examinons l'effet d'annonce des news relatifs aux politiques monétaires de la BCE et de la FED et issus des réunions officielles du Conseil des gouverneurs et du FOMC sur la volatilité intrajournalière du taux de change euro-dollar à cinq minutes d'intervalles. Les résultats montrent que les news de la politique monétaire de la BCE relatifs à ses taux d'intérêt Target sont plus significatifs et plus influents sur le niveau de la volatilité intrajournalière que ceux de la politique monétaire de la FED relatifs à son taux des fonds fédéraux. Malgré le nombre réduit de ces news, leur effet apparaît statistiquement significatif au cours des années de l'échantillon du taux de change euro-dollar choisi. Nous avons également introduit une structure polynomiale qui nous permet de prendre en compte la persistance de court terme et de mettre en évidence une possible dissymétrie dans l'effet de chaque variable de signal sur la volatilité du taux de change euro-dollar.

Evaluation des Densités de Prévision : Une Approche GMM, *Jaouad Madkour*

Nous proposons deux tests d'évaluation des densités de prévisions fondés sur la Méthode des Moments Généralisée, l'un pour tester la loi Uniforme U[0,1] à l'instar de Diebold et al. (1998) et l'autre pour tester la loi Normale centrée et réduite de la même manière que Berkowitz (1999). Outre leur simplicité et leur souplesse, des expériences Monte Carlo ont mis en évidence leur fiabilité et leur puissance en particulier sur des échantillons de taille réduite. Leur utilisation dans la validation des densités de prévisions du taux de change, à côté du test du ratio de vraisemblance de Berkowitz (1999), a confirmé des résultats déjà trouvés dans la littérature, à savoir, que le modèle SETAR est le plus à même à reproduire les non linéarités observées dans ces séries et, par conséquent, à fournir des prévisions précises.

Testing Interval Forecasts : A New GMM-Based Test, *Jaouad Madkour, Christophe Hurlin and Elena-Ivona Dumitrescu*

Nous proposons un cadre unifié pour évaluer les prévisions par intervalle de confiance fondé sur les travaux de Bontemps and Meddahi (2005,2006). En testant l'hypothèse que la série des

violations est distribuée selon une loi de Bernoulli, nous montrons que notre test, basé sur la méthode des moments généralisée, des hypothèses de couverture non conditionnelle, d'indépendance et de couverture conditionnelle respectivement dépasse le test du ratio de vraisemblance de Christoffersen (1998) sur des échantillons de taille similaire à celle des séries disponibles. De telles conclusions sont confirmées par l'application aux indices SP500 et Nikkei.

Eliciting Individual Preferences for Pension Reform using a Discrete Choice Experiment (Evaluer les Préféreces Individuelles pour la Réforme des Retraites à travers la méthode des Choix d'Options Répétées), *Yosr Abid and Cathal O'Donoghue*

Les systèmes de retraite ont récemment été scrupuleusement examinés en raison du vieillissement de la population prévu menaçant leur soutenabilité. La contribution de ce papier au débat se fait d'une perspective politico-économique puisqu'il utilise des données d'une enquête de préférences établies du type choix d'options répétées (choice experiment) afin d'investiguer les préférences individuelles pour un plan de retraite public alternatif. Les résultats suggèrent que le système de valeur des individus est un déterminant important de leurs préférences. Le revenu détermine le degré de redistribution préféré. Cependant, les préférences en fonction de l'âge sont en contradiction avec ce qui est suggéré en théorie. A notre connaissance, ceci représente la première tentative d'utilisation d'une telle méthodologie dans le domaine des préférences pour la redistribution à travers les systèmes de retraite.

Le classement et le repérage socioéconomique du ménage : cas des ménages algériens, *Mahali Kamel*

Un des outils d'observation du social est d'une part la nomenclature des catégories socio professionnelles si l'étude porte sur l'individu et d'autre part les catégories socio économiques dans le cas du ménage. Ces deux outils permettent de fournir un découpage de la société selon des critères à la fois économiques et sociaux. Classer les ménages dans des groupes sociaux revient d'un coté à définir les indicateurs et les critères de classification, et d'un autre coté avoir des informations bien précises et fiables sur les éléments à classer. Ces informations concernent les individus constituant le ménage (niveau d'instruction, catégorie socioprofessionnelle, etc.) et le ménage dans ça totalité (nombre de personne par ménage, revenu du ménage, etc.). Donc la construction de la typologie (groupes sociaux) doit passer d'un niveau individuel (Catégories socioprofessionnelles) à un niveau collectif (Catégories socio-économiques). Ainsi le point central de cette étude est de développer la réflexion sur le passage des individus comme entités aux ménages comme autres entités à travers les pratiques algériennes.

Problème de Bandit Unimodal, *Jia Yuan Yu*

L'hypothèse d'unimodalité présente un avantage important : nous pouvons déterminer si un bras est optimale en échantillonant les directions autour de celui-ci. Cette propriété nous permet de trouver plus efficacement le bras optimale, ainsi que de détecter plus rapidement des

changements abrupts dans les distributions des récompenses. Par example, dans le cas du bandit sur un graphe, l'écart de performance est proportionnel, non au nombre total de sommets, mais au degré maximal et au diamètre du graphe.

Extension du modèle RLAR (Random Level Shift Autoregression) : Estimation bayesienne et modélisation du prix du baril de pétrole, *Oumelkheir Moussi*

Cet article a pour but l estimation des paramètres d un modèle dérivé du modèle autorégressif avec rupture en moyenne(Randomlevel shift model),suivi d une application sur une série de prix du baril de pétrole.La méthode d estimation utilisée est la méthode bayesienne avec application de l échantillonnage de Gibbs. Mots clés : échantillonnage de Gibbs, densité à priori, densité à postériori, prix du baril de pétrole.

Renforcement de la chaîne logistique par l'entreposage de données trajectoires, *Nakhla Zina and Akaichi Jallel*

Pour améliorer la rentabilité et la qualité de services, et faire face à une concurrence de plus en plus ardue, les entreprises sont à la recherche des approches efficaces permettant d'améliorer leurs métiers en générales, et la gestion de la chaîne logistique en particulier qui joue un rôle primordiale dans la réduction des couts, l'amélioration de la qualité des services, et l'augmentation de la productivité. Ce travail s'inscrit dans le cadre d'amélioration de la performance d'une chaîne logistique par la conception d'un Entrepôt de Données Trajectoires (EDT) destiné à collecter les données relatives aux objets mobiles. Les données de l'entrepôt seront analysées pour extraire des connaissances servant à une prise de décision conduisant à renforcer la gestion de la chaîne logistique.

Inférence asymptotique dans les processus ARCH(q) périodiques, *Ines Lescheb*

Dans la suite de ce papier on va étudier les propriétés asymptotiques et l'estimation des paramètres pour les processus ARCH périodiquement corrélés (PC). On donne la définition d'un processus PARCH et quelques propriétés probabilistes en basant sur les représentations BEKK et VEC. Aussi j'ai établi les propriétés asymptotiques de la moyenne empirique et la covariance.

An optimal confidence interval for an adjusted premium estimator in simulated insurance data, *Kmar Fersi, Kamel Boukhetala and Samir Ben Ammou*

The aim of this work is to determine an optimal confidence interval for the adjusted premium estimator, developed by [7] in non-life insurance. For this, we are led to improve the minimization variance strategy of this estimator under stochastic constraints related to excesses over

thresholds u ([4]). The quality of the results of this stochastic minimization problem, obtained by Genetic Simulated annealing Algorithm (GSA) ([6] and [10]), is very sensitive to the number of extreme risks. When the proportion of the excess is low, a simulation technique based on actual observations is proposed to determine an adequate size of simulated sample likely to calculate an optimal confidence interval for the estimator.

Testing Scale Efficiency : A Smooth Bootstrap Approach, *Hedi Essid, Pierre Ouellette and Stéphane Vigeant*

This paper presents a nonparametric statistical test procedure for organization scale efficiency. This procedure allows the practitioner to test whether the observed scale efficiency score is real or due to sampling variation. We use the Data Envelopment Analysis method (DEA) to estimate efficiency scores and we adopt a smooth homogeneous bootstrap methodology to approximate the sampling distribution of the test statistic. The results of Monte Carlo simulation show the performance of our testing procedure.

PCA, FA, ICA and LDA algorithms for Data reduction, Discriminent analysis, Classification and Knowledge extraction of complex biological data, *Ali Mohammad-Djafari and Ghazale Khodabandelou*

In this paper, first we present, in a unifying way, Principal Component Analysis (PCA), Factor Analysis(FA), Independent Component Analysis (ICA) and Linear and Nonlinear Discriminent Analysis (DA) methods and associated practical algorithms which can be used in Data Reduction (DR) and supervised classification of multivariate great dimentional data. Then, we present A Matlab toolbox which gives the possibility to simulate and test these algorithmes and show simulation results which show the performances of these algorithms. In the second part, we describe some biological experiments related to studying the circadien cell cycles of mouses and cancer treatement : In a first step to observe different kind of data (Tempreature, Activity, Hormons, Genes and Proteins) to understand the complex biological and medical effects. The biologists need to visualize, to analyse and to do classifications on these data and finally to extract some knowledge from them and propose biological models describing the studied phenomena. These data are often complex : multivariate, great dimensionality, heterogeneous, with missing data, and observed at different sampling rates. The classical methods of PCA, FA, ICA and LDA can not directly handle these data. In this paper, first we show the performances of these algorithms on real data and then propose prospective new extentions to push further their limits.

Improved Dynamic Weighted Majority algorithm for parameter selection, *Dhouha Mejri and Mohamed Limam*

Dynamic weighted majority-Winnow (DWM-WIN) algorithm is a powerful classication method that handles nonstationary environment and copes with concept drifting data streams. Although its good performance, this method has a serious problem in choosing the best values

of the algorithm parameters that affects the classication performance of DWM-WIN. Hence, there is a need for a rational automatic selection of parameter values. To deal with this issue, a genetic algorithm (GA) is used as an optimization method to find the best values of these parameters. We have used DWM-WIN as a fitness function of GA and a set of several parameters combination as initial solutions for GA. In order to assess this optimized DWM-WIN algorithm, four data sets are simulated from UCI data sets repository to highlight the effectiveness of the optimized DWM-WIN compared to other algorithms.

Comparaison de taux d'incidence par des modèles de régression dérivés de Poisson, *Sandrine Domecq, Marion Kret, Christelle Minodier and Philippe Michel*

Les taux d'incidence des événements indésirables graves observés au cours de deux enquêtes prospectives en milieu hospitalier menées en 2004 et en 2009 ont été comparés. Le nombre d'événements indésirables graves associés aux soins, observés au sein d'unités d'hospitalisation sur 7 jours maximum, a été considéré comme la réalisation d'une variable aléatoire discrète suivant une loi de Poisson (2001). Partant de cette hypothèse, les taux d'incidence ont été comparés en utilisant des modèles de régression de Poisson ou binomials négatifs en cas de surdispersion (2009). La surdispersion a été vérifiée grâce au test de Dean. Le nombre de jours d'observation a été pris en compte dans un terme offset. Des caractéristiques au niveau des unités d'hospitalisation ont été intégrées dans les modèles comme variables d'ajustement. De nombreuses comparaisons ont été faites sur des sous-échantillons d'événements indésirables graves. L'analyse a porté sur 8754 séjours hospitaliers représentant 35234 jours d'observation dans 294 unités d'hospitalisation en 2004 et sur 8269 séjours représentant 31663 jours d'observation dans 251 unités d'hospitalisation en 2009 (2010). Ainsi, l'analyse multivariée a permis d'interpréter les différences entre taux d'incidence en termes de risques relatifs en tenant compte de plusieurs variables d'ajustement et du temps d'exposition. Cette application illustre l'intérêt d'utiliser des modèles de régression dérivés de Poisson par rapport aux méthodes de standardisation pour comparer des taux d'incidence (1995).

Conception de chimiothèques enrichies en inhibiteurs d'interactions protéine-protéine, *Christelle Reynes, Anne-Claude Camproux, Bruno Villoutreix and Olivier Sperandio*

Les interactions protéine-protéine (IPP) sont susceptibles de devenir une catégorie majeure de cibles thérapeutiques. Actuellement, une infime partie des molécules thérapeutiques disponibles cible ce type d'interaction, alors que la taille de l'interactome humain semble être très importante (650,000 interactions estimées). Les caractéristiques biologiques de ces systèmes empêchent d'utiliser efficacement les filtrages traditionnels pour molécules thérapeutiques. Il y a donc un réel intérêt à mieux comprendre l'espace chimique recouvert par les inhibiteurs d'IPP pour pouvoir concevoir des chimiothèques de composants orientées vers ce type de chimie. Des arbres de classification ont été utilisés pour construire un modèle distinguant les inhibiteurs de PPI d'autres types de molécules thérapeutiques. Le modèle obtenu a mis en évidence deux

descripteurs reflétant des formes moléculaires spécifiques ainsi que la présence d'un certain nombre de liaisons aromatiques. Deux versions de ce modèle ont été implémentées dans un programme gratuit (PPI-HitProfiler) permettant de savoir si une molécule est susceptible d'être un inhibiteur de PPI.

Classification des données quantitatives de grande dimension dans l'environnement logiciel mixmod, *Christophe Biernacki, Gilles Celeux, Gérard Govaert and Florent Langrognet*

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).

Elaboration d'un âge biologique à partir de données accessibles en routine de médecine généraliste : Essai de fondement théorique, *Marianne Sarazin*

Le vieillissement caractérise une évolution inéluctable du corps dont la quantification est établie par l'âge dépendant du temps dit "'chronologique"'. Cependant, ce critère âge ne quantifie qu'imparfaitement l'usure réelle du corps soumise à de nombreux facteurs modificateurs dépendant des individus. Aussi, a-t-il été substitué depuis longtemps par un critère composite, appelé "'âge biologique"', sensé davantage refléter le vieillissement individuel. Afin d'essayer d'en faire un outil quantificateur accessible à la pratique de médecine générale, une nouvelle méthodologie est proposée. Le critère "'âge biologique"' a été défini à partir de la grandeur âge "'chronologique"' et de variables cliniques et biologiques pondératrices caractérisant l'état de santé du corps humain et mesurables au cours d'un examen médical standard. Un échantillon de population témoin supposé "'vieillir normalement"', selon les critères de normalité des variables utilisées, a servi pour le calcul du rôle pondérateur de chacune des variables. Le sexe a été au préalable fixé. La dépendance statistique des variables utilisées a été modélisée par une copule gaussienne (prise en compte seulement de corrélations linéaires deux à deux). L' "'âge biologique"' standardisé a alors été défini explicitement à partir des coefficients de corrélation ainsi calculés. La validité théorique du modèle a été prouvée lorsque l'âge chronologique était entièrement expliqué par les variables biologiques. Par ailleurs, pour chaque sous-échantillon,

les queues de distribution des lois marginales ont été estimées (méthode des excès) afin de renforcer le pouvoir discriminant du modèle. Cette méthode ouvre de nouvelles perspectives en termes de prévention et prise en charge du risque de vieillissement. Cependant, la pertinence de ce critère doit être validée par des études de morbidité et un retour d'expérience des médecins généralistes.

L'approche PLS pour la recherche de marqueurs dans le cadre d'une étude clinique observationnelle en nutrition, *Marie Keravec and Pascale Rondeau*

Lors d'une étude clinique observationnelle en nutrition, de nombreuses données de natures différentes sont collectées : recueils alimentaires, données provenant de différentes matrices biologiques. L'un des objectifs de ce type d'étude est d'identifier des biomarqueurs liés à la consommation alimentaire.

Afin de répondre à cet objectif, l'approche PLS est évaluée sur une matrice de données comprenant quatre-vingt-deux sujets, quatre-vingt-deux paramètres biologiques et la consommation journalière de fluides. Les sujets sont répartis en trois catégories de consommation : trente-neuf petits buveurs (sujets déclarant des consommations de fluides inférieures à 1,2L/ jour), onze buveurs moyens (consommations entre 1,2 et 2L/jour) et trente-deux gros buveurs (consommations supérieures à 2L/jour).

Un modèle PLS1 conduit à l'identification d'une dizaine de marqueurs urinaires cependant, l'erreur d'estimation du modèle est élevée ($\pm$597 mL) par rapport à la quantité moyenne de fluide déclarée par jour (environ 1600 mL). Cette imprécision d'estimation peut être due à la grande variabilité de la consommation de fluides Y par rapport à une faible variabilité des marqueurs biologiques sélectionnés *(Xj)*.

Au vu de ces résultats, une PLS DA a été menée afin d'évaluer la robustesse des marqueurs identifiés. Afin de s'affranchir de l'hypothèse de linéarité du modèle PLS, la méthode de classification supervisée random forest a également été appliquée. Les résultats issus de ces méthodes seront présentés et discutés.

On stationarity and existence of moments of the spatial RCA models, *Karima Kimouche*

In the present paper, we give necessary and sufficient conditions ensuring the stationarity and the existence of higher-order the moments of a spatial autoregressive random coefficient models. The spectral density function of the process is obtained.

Statistical properties of Parasite Density estimators in Malaria and field applications, *Imen Hammami, Grégory Nuel and André Garcia*

Malaria is a global health problem responsible for nearly 3 million deaths each year, an average of one person every 12s. In addition, 300 to 500 million people contract the disease each year. The level of infection, expressed as the parasite density (PD), is classically defined

as the number of asexual forms of Plasmodium falciparum relative to a microliter of blood. Microscopy of Giemsa-stained thick blood films is the gold standard for parasite enumeration in case of febrile episodes. PD estimation methods usually involve threshold values as the number of white blood cells (WBC) counted and the number high power fields (HPF) seen. However, the statistical properties of PD estimates generated by these methods have been generally overlooked. Here, we study the statistical properties (bias, variance, False-Positive Rates...) of the PD estimates of two commonly used threshold-based counting techniques according to varying threshold values. Furthermore, we give more insights on the behavior of measurement errors according to varying threshold values and on what would be the optimal threshold values that minimize the variability.

Méthodologie d'inversion des mesures optiques (AOT) en mesures de qualité de l'air (PM10) basée sur les réseaux de neurones, *Houda Yahi, Sylvie Thiria and Michel Crepon*

Nous présentons une méthode d'inversion de mesures photométriques d'épaisseurs optiques (AOT) en concentrations massiques atmosphériques (mesures de qualité de l'air PM10). Comme les PM10 est une mesures de surface et l'AOT est une mesure intégrée, la relation liant les deux mesures est très complexe. Étant donné que ces deux paramètres dépendent fortement de structures atmosphériques et des paramètres météorologiques, nous avons classé les situations météorologiques en termes de types de temps en utilisant un classificateur neuronal (cartes auto-organisatrices). Pour chaque type de temps, nous avons constaté que la relation entre l'AOT et de PM10 peut être établie et l'inversion effectuée à l'aide de réseaux de neurones avec des performances satisfaisantes. Afin d'accroître la fiabilité statistique de la méthode nous avons appliqué cette approche à la région de Lille (France) pour la période de cinq d'été (étés des années 2003-2007).

,

Modélisation Multi-variée des extrêmes hydrométéorologiques-Application : Mildiou de la vigne, *Dhouha Ouali, Zoubeida Bargaoui and Samir Chbil*

La notion du risque est une notion très complexe ; Elle découle de la conjonction d'un aléa non maîtrisé et de l'existence d'un enjeu ou d'un environnement pouvant être affecté par un tel événement. Ainsi, nous visons par la présente étude développer un processus d'identification du risque, en se basant sur la théorie statistique des valeurs extrêmes. En fait, cette dernière présente un outil adéquat de maitrise et d'aide à la décision via la modélisation de l'occurrence des événements extrêmes. Notre étude consiste à étudier le risque lié au Mildiou de la vigne dans les vignobles du Cap Bon-Tunisie, et à proposer une approche de précaution afin d'obtenir les objectifs fixés en termes de qualité et de quantité avec un minimum d'interventions. Les variables à modéliser sont la température et les précipitations ; Du fait de la structure de corrélation de ces deux variables, une analyse fréquentielle univariée ne permettra pas de bien évaluer les probabilités au dépassement, ainsi nous avons eu recours à la théorie des copules

pour effectuer une modélisation bivariée de ces deux variables.

Analyse de données d'expression des gènes impliqués dans la polyarthrite rhumatoïde, *Sonia Kechaou-Cherif, Slimane Ben Miled and Alia Benkahla*

La polyarthrite rhumatoïde est un rhumatisme inflammatoire chronique. C'est une maladie auto immune d'étiologie incertaine caractérisée par l'inflammation chronique et symétrique des grandes et des petites articulations. Il s'agit du rhumatisme inflammatoire le plus fréquent. Il affecte environ 1Le mécanisme physiopathologique de cette maladie n'est pas encore élucidé et on pense que plusieurs gènes dont l'expression change pourraient être à l'origine de cette maladie. L'objectif de notre étude est d'identifier ces gènes et de montrer via une méta-analyse leur implication dans la polyarthrite rhumatoïde. Pour cela, nous avons prélevé des données sur des bases de données publiques "GEO" et "Array Express". Nous avons sélectionnées des données brutes que nous avons normalisées puis corrigées. Nous avons procédé par la suite à une analyse de variance en utilisant la méthode ANOVA afin de déterminer les gènes différentiellement exprimés. Nous avons également procédé à un classement hiérarchique de tous les gènes par la méthode dite "kmean".

Lecture Probabiliste du Cycle Boursier Tunisien : Proposition d'un modèle à trois états avec changements de régimes markoviens et dépendance à la durée, *Adel Karaa, Emna Mahat and Azza Bejaoui*

Les méthodes linéaires existantes qui visent généralement à reproduire statistiquement les dynamiques boursières échouent à mettre en lumière les principales propriétés des indices boursiers. Une bonne compréhension de ces caractéristiques demeure un enjeu scientifique majeur pour les économètres et financiers afin de mieux appréhender le comportement des rendements boursiers. Pour cela, le modèle à changement de régime markovien, développé dans le cadre de l'étude de l'indice boursier TUNINDEX, introduit la notion de la dépendance à la durée afin de mieux étudier les dynamiques boursières régissant le marché boursier tunisien ainsi capter sa structure non-linéaire. Le modèle proposé caractérise les trois états trouvés comme des états haussier, normal et baissier. La matrice de transition sous-jacente à ce modèle admet une forme très particulière qui offre la possibilité d'étudier de près le comportement des investisseurs sur le marché boursier tunisien. Par ailleurs, le trade-off rendement-risque n'est pas stable à travers les états ce qui renvoie à l'hétérogénéité du comportement des investisseurs durant chaque état. Finalement, notre modèle à 3 états offre des implications managériales intéressantes en termes de gestion du risque et de décisions d'investissement.

Vers une approche probabiliste de la dépendance à la durée et de la datation du cycle boursier tunisien, *Emira Torjmen and Adel Karaa*

A travers cet article, nous proposons d'identifier et d'analyser les différentes phases cycliques qui caractérisent la dynamique du marché boursier tunisien en utilisant un modèle à changements de régimes markovien. Notre étude porte sur une série de rentabilités hebdomadaires

de l'indice boursier TUNINDEX sur une période s'étalant du 07/01/1998 au 19/08/2010. La modélisation de cette série est représentée par un modèle MS-AR pour les rendements et un modèle MS-EGARCH pour la volatilité. L'analyse des données nous a conduit à accepter le caractère non linéaire du processus sous-jacent à la série des rentabilités. Les résultats obtenus mettent en évidence la présence de trois régimes distincts relatifs aux types des événements annoncés sur le marché , à savoir ; un régime hors événements, un régime événementiel de bonnes nouvelles et un régime événementiel de mauvaises nouvelles. De même, les probabilités obtenues dans le cadre de notre étude montrent une forte persistance du régime hors événements et une forte dépendance à la durée des différents régimes. Dans cet article, nous fournissons aussi une datation des fluctuations cycliques des rendements de l'indice.

Estimation bayésienne d'un modèle de volatilité stochastique, *Chouik Belmokhtar and Anes Ouali*

Dans les modèles ARCH, introduits par Engel (1982), la volatilité est considérée comme une fonction déterministe. L'autre alternative est fournie par les modèles de volatilité stochastique (SV). Cette classe (SV), introduite par Taylor (1982), considère que la volatilité est un processus stochastique latente et offre plus de flexibilité dans la modélisation des données. Les modèles (SV) sont plus difficiles à estimer que les modèles traditionnels de type ARCH. Jacquier E et all (1994) ont considéré un modèle (SV) qui possède trois paramètres, le facteur d'échelle de la volatilité, le paramètre de persistance des chocs et la volatilité de la log-volatilité. Plusieurs approches ont été utilisées pour l'inférence statistique des paramètres du modèle (SV), nous citons la méthode des moments abordée par Taylor (1986) et l'approche bayesienne utilisée par Jacquier E et all (1994), (2004). Nous reprenons l'approche Bayesienne pour une inférence statistique de ces paramètres. Sur la base d'une loi a priori gamma-normale pour les paramètres du modèle, les estimations bayésiennes ne peuvent pas être obtenues analytiquement mais les méthodes de Monte Carlo (MCMC) nous permettent d'approcher ces estimations . Les distributions conditionnelles des paramètres, étant usuelles , nous pouvons simuler selon ses lois, ce qui n'est pas le cas pour celle des volatilités. Nous proposons une augmentation de données pour se permettre d'appliquer le procédé d'échantillonnage de Gibbs et donc simuler selon les distributions conditionnelles des volatilités. A l'aide du logiciel R, nous générons un processus selon un modèle SV et nous appliquons les méthodes (MCMC) basées sur les techniques d'échantillonnage de Gibbs. Enfin nous notons que les valeurs estimées de l'espérance et la variance a posteriori des paramètres du modèle sont proches des vraies valeurs associées.

Announcement effect and intraday volatility patterns of euro-dollar exchange rate : monetary policy news arrivals and short-run dynamic response.

Mokhtar DARMOUL*, Mokhtar KOUKI†

Résumé

Dans cet article, nous examinons l'effet d'annonce des news relatifs aux politiques monétaires de la BCE et de la FED et issus des réunions officielles du Conseil des gouverneurs et du FOMC sur la volatilité intrajournalière du taux de change euro-dollar à cinq minutes d'intervalles. Les résultats montrent que les news de la politique monétaire de la BCE relatifs à ses taux d'intérêt Target sont plus significatifs et plus influents sur le niveau de la volatilité intrajournalière que ceux de la politique monétaire de la FED relatifs à son taux des fonds fédéraux. Malgré le nombre réduit de ces news, leur effet apparaît statistiquement significatif au cours des années de l'échantillon du taux de change euro-dollar choisi. Nous avons également introduit une structure polynomiale qui nous permet de prendre en compte la persistance de court terme et de mettre en évidence une possible dissymétrie dans l'effet de chaque variable de signal sur la volatilité du taux de change euro-dollar.

Abstract

In this article, we examine the announcement effect of news relating to the monetary policies of the ECB and the FED and resulting from the official meetings of the Council of the governors and the FOMC on intraday volatility of the foreign exchange rate euro-dollar at five minutes of intervals. The results show that the news of the monetary policy of the ECB relative to its Target interest rates are more significant and more influential on the level of intraday volatility than those of the monetary policy of the FED relative to its federal funds rate. In spite of the reduced number of these news, their effect appears statistically significant during the years of the sample of foreign exchange rate euro-dollar selected. We also introduced a polynomial structure which enables us to take into account the short-run response patterns and to highlight a possible dissymmetry in the effect of each variable of signal on the volatility of foreign exchange rate euro-dollar.

Mots clefs : Effet d'annonce, Forex, News, Taux de change.

Classification JEL : C15, E44, F31, G14

*Université Paris 1 La Sorbonne (TEAM)

†LEGI-Ecole Polytechnique de Tunisie, Ecole Supérieure de la Statistique et de l'Analyse de l'Information - Tunis.

1 Introduction

Un grand nombre de travaux dans la littérature économique et celle financière traitent de l'effet des *news* économiques sur les rendements des actifs financiers, mais de manière apparemment différente. Tout d'abord, alors que les économistes ont tendance à se concentrer sur l'impact des *news* sur le niveau des rendements des actifs tels les travaux de Bomfim et Reinhart (2000), Kuttner (1999), Roley et Sellon (1998), Thornton (1998), et Reinhart et Simin (1997), leurs homologues financiers comme Andersen et Bollerslev (1998), Jones, Lamont, et Lumsdaine (1998), Berry et Howe (1994), Mitchell et Mulherin (1994), Ederington et Lee (1993), et Cutler, Poterba, et Summers (1989) ont, par contre, cherché à trouver un rapport entre ces *news* et la volatilité des rendements. En second lieu, bien que la majorité des travaux dans la littérature économique n'a pas bien pris en compte l'effet de surprise des *news* de politique monétaire et leur impact sur les différents marchés financiers, peu d'études dans la littérature financière se sont concentrées réellement sur l'impact des *news* de politique monétaire sur la volatilité des marchés. En outre, les travaux mettant en relation *news*-volatilité n'ont, généralement, pas fait de distinction entre les annonces officielles et celles non-officielles.

Dans cet article, nous essayons d'établir un lien entre les deux littératures en analysant les effets d'annonce (effets des informations publiques) qui sont différents des autres composantes de la volatilité des actifs financiers tel que l'effet calendrier. Ces effets ressemblent à des processus d'ajustement instantanés des prix qui induisent des éclats de volatilité importants, mais de courte durée. Nous étudions, également, l'impact de toute la grille des annonces officielles programmées des politiques monétaires de la Banque Centrale Européenne (BCE) et de la *Federal Reserve Bank* (FED), visant leurs taux d'intérêt *Target* (Moschitz (2004)). Nous concentrons notre recherche sur la signification de chaque type d'annonce et le comportement de la dynamique de réponse de la volatilité associée. Ceci nous paraît approprié pour modéliser le processus de la volatilité du taux de change euro-dollar, mais également intéressant comme une mesure de la significativité de chaque type d'annonce.

Notre étude peut constituer, de différentes manières, un enrichissement de la littérature disponible en la matière. En particulier, l'utilisation des cotations *spot* continues du marché, sur les 24 heures du cycle des transactions du marché FOREX, nous permet d'étudier l'effet de tous les signaux des annonces officielles de politique monétaire concernant les taux d'intérêt *Target* de l'Europe et des États-Unis, dont l'influence sur le comportement de la volatilité intrajournalière du taux de change euro-dollar n'a pas été analysée antérieurement dans la littérature. D'autre part, contrairement à l'analyse d'Ederingtone et Lee (1993) qui se base sur les prix des titres aux États-Unis, nous pouvons affirmer qu'une fois la variabilité intrajournalière systématique est correctement expliquée, les effets d'annonce, bien que statistiquement significatifs, sont d'une grande importance pour l'explication de la volatilité globale. En effet, les annonces officielles exercent clairement une influence dominante sur quelques intervalles prévisibles de cinq-minutes juste après la communication des *news*.

Nous allons chercher, dans cet article, à déterminer si le pouvoir explicatif de ces annonces (effet annonce) est meilleur ou non que celui relatif au comportement intrajournalier de la volatilité (effet

calendrier).

Notre travail sera organisé comme suit : dans la première section nous allons introduire le type de *news* ainsi que l'échantillon de taux de change utilisés et, ensuite, analyser le comportement de la volatilité intrajournalière du taux de change euro-dollar en distinguant l'effet de chaque variable de *news*. Dans la deuxième section, nous allons présenter une structure polynomiale qui pourrait, dans de futurs travaux, être très utile pour la prise en compte et la modélisation de la persistance de court terme de l'effet d'annonce dans la volatilité conditionnelle des taux de change à très haute fréquence.

2 Le comportement de la volatilité sous l'effet d'annonce :

Nous avons décomposé les annonces relatives aux politiques monétaires visant les taux d'intérêt Target de la BCE et de la FED en quatre catégories distinctes, nous permettant, en premier lieu, d'observer l'influence de chaque type d'annonce sur le comportement de la volatilité intrajournalière et, en second lieu, de montrer que l'influence de cet effet d'annonce est aussi importante que celle de l'effet calendrier dans le comportement de la volatilité.

Dans cette section, nous allons étudier l'influence de chaque type de *news* de politique monétaire, séparèment, en rappelant leurs notations respectives :

- *bce-rv* représente les *news* émanant du résultat de la réunion du Conseil des gouverneurs, c'est à dire l'information lancée aux opérateurs du marché FOREX sur les décisions de la BCE annonçant des variations concernant les taux d'intérêt directeurs (le taux de soumission minimal appliqué aux opérations principales de refinancement ainsi que les taux d'intérêt de la facilité de prêt marginal et de la facilité de dépôt) ;
- *bce-rnv* représente les *news* donnant des informations sur la politique de maintien de ces taux à des niveaux fixes ;
- *fed-rv* représente les *news* informant le marché des décisions prises par le FOMC concernant l'orientation de la politique monétaire américaine et révélant une variation au niveau du taux des fonds fédéraux de la FED ;
- et *fed-rnv* représente les *news* renseignant le marché sur la stabilité de la politique monétaire américaine, c'est à dire ne révélant pas de variations de ce taux d'intérêt.

3 L'évaluation des effets d'annonce dans le comportement de la volatilité intrajournalière :

3.1 L'impact des signaux de politique monétaire sur la volatilité du taux de

change €/$:

Il est admis, dans ce travail, que le marché est totalement efficient. De ce fait, les prix des actifs devraient refléter toute l'information disponible, et les variations de taux de change sont fortement influencées par l'arrivée d'une nouvelle information sur le marché FOREX. Ceci suppose, toutefois, que l'information n'a pas été anticipée, sinon elle serait déjà intégrée dans le taux de change. Nous appelons, dans ce qui suit, *news* « signal politique » cette nouvelle information officielle non anticipée de l'évolution des taux d'intérêt cibles sur le marché monétaire, laquelle information se transmet au marché FOREX à travers les taux de change. Si la politique monétaire est un déterminant important du taux de change, les *news* de politique monétaire doivent avoir un impact significatif sur le niveau du taux de change.

Un signal politique qui a été « parfaitement » anticipé par les agents n'affecte pas le niveau du taux de change, mais il peut avoir un impact sur la volatilité du taux de change. En se référant à la théorie du signal introduite par Mussa (1981) et utilisée par Dominguez et Frankel (1993b), Dominguez (2003), qui repose sur l'hypothèse d'asymétrie de l'information (les autorités monétaires possèdent une information quant à la politique monétaire future supérieure à celle détenue par le marché), l'effet des signaux de politique monétaire envoyés par les banques centrales (au moyen de leurs réunions) sur la volatilité va dépendre de la manière dont les agents les perçoivent. Si ces signaux sont jugés parfaitement crédibles et non ambigus, ils devraient soit ne pas influencer la variance conditionnelle du taux de change soit réduire la volatilité. Par contre, si ces signaux sont perçus par le marché peu crédibles ou confus, ils devraient accroître l'incertitude et par là même la volatilité, et leur impact serait plus ou moins important selon la précision de l'information révélée par le signal.

Dans cet article, nous distinguons deux catégories de signaux issus des réunions officielles des deux banques centrales, ceux révélant des variations des taux d'intérêt et ceux reportant des taux d'intérêt inchangés. Normalement, il sera attendu à ce que l'effet sur la volatilité des signaux annonçant une variation des taux d'intérêt soit plus élevé que celui des signaux n'annonçant pas de variation. En effet, l'annonce précise d'une variation du taux d'intérêt fournit aux opérateurs un meilleur point focal "*Benchmark*" pour qu'ils mettent à jour leurs opinions, par rapport au cas où l'annonce n'indique pas de variation. Cette annonce de variation du taux d'intérêt peut être, pour les opérateurs, une confirmation d'une action de la Banque Centrale qui n'était pas certaine et peut provoquer une révision des anticipations et, donc, une variation de leurs positions. Les opérateurs vont, alors, échanger pour atteindre leurs objectifs, et ces échanges génèrent de la volatilité. Par contre, un signal qui ne mentionne aucune variation du taux d'intérêt peut rendre la convergence des opinions uniquement partielle. Ainsi, la révision des anticipations ne peut être que partielle, le montant des échanges sera plus faible et, par conséquent, la volatilité plus basse.

Nous postulons dans cette analyse une relation positive entre le montant des transactions et la volatilité. Toutefois, dans la littérature sur les microstructures, cette relation peut être plus complexe. Ainsi, pour Jorion (1996), autant au plan théorique qu'à partir d'études économétriques, la corrélation

est effectivement positive en présence d'anticipations fortement hétérogènes, mais devient négative si les anticipations sont convergentes ; le marché gagnant en résilience quand le nombre d'opérateurs et les volumes d'échanges augmentent.

3.2 La persistance de court terme des signaux :

Pour examiner l'effet de persistance des signaux de la BCE et de la FED sur la volatilité du taux de change €/\$, il nous faut extraire leurs effets restants dans la volatilité (après avoir filtré les données de l'effet calendrier). Cette évaluation de la persistance de l'effet d'annonce nous a paru très compliquée, en raison du fait que les coefficients de régression de ce type d'effet ne sont pas des indicateurs simples, mais impliquent des comportements pré-spécifiés de la dynamique de réponse de la volatilité intrajournalière. Il aurait pu être possible d'utiliser une matrice avec une spécification simple consistant à introduire, dans l'équation de la volatilité, des variables binaires prenant la valeur 1 au moment du signal et 0 sinon pour chaque signal. En particulier, si nous supposons que les k différents types de signaux auraient un impact significatif sur la volatilité au cours des N_k intervalles de cinq minutes, alors nous imposons une structure raisonnable au comportement de réponse de la volatilité intrajournalière, en estimant tout simplement le degré auquel l'événement persiste dans ce comportement. Pour cette raison, contrairement à Andersen et Bollerslev (1998), nous proposons d'approcher la structure de persistance du $k^{ième}$ signal au moyen d'un polynôme d'ordre h :

$$\Gamma_k(i) = \sum_{j=0}^{h-1} \gamma_{j,k} \left[1 - \left(\frac{i}{N_k} \right)^{h-j} \right] .i^j \tag{1}$$

avec $i = 0, 1, \ldots, N_k$ et $j = 0, 1, \ldots, h-1$.

Par conséquent, chaque séquence (vecteur) de $\Gamma_k(i)$ correspond à l'impact du $k^{ième}$ signal à l'horizon i et $\gamma_{0,k}(0)$ est l'effet instantané.

Cependant, nous intégrons dans ce polynôme une matrice I_k de variables indicatrices "*dummies*". Ces variables indicatrices sont relatives à chaque type de signal k (les quatres types de *news* retenus). De ce fait, notre fonction d'impact peut s'écrire :

$$\sum_{k=1}^{K} \sum_{i=0}^{N_k} \Gamma_k(i) .I_k(t, n-i) \tag{2}$$

avec $k = 1, 2, \ldots, K$, $(K = 4)$, $n = 1, 2, \ldots, N$, $(N = 288)$, $t = 1, 2, \ldots, T$, $(T = 782)$, $i = 0, 1, \ldots, N_k$, $(N_k = 12)$, et $I_k(t, n-i)$ est un indicateur d'occurrence matriciel des *dummies* du $k^{ième}$ signal durant le $n^{ième}$ intervalle du jour t.

Références

[1] Andersen T. Bollerslev T. (1998), " DM-Dollar Volatility : Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies ", The Journal Of Finance, 53(1), pp. 219-265.

[2] Berry D. Howe M. (1994), " Public Information Arrival ", Journal of Finance, 49, pp.1331-1346.

[3] Bomfim N. Reinhart R. (2000), " Making News : Financial Market Effects of Federal Reserve Disclosure Practices ", Manuscript, Federal Reserve Board.

[4] Cutler M. Poterba M. Summers H. (1989), " What moves stock prices ? ", Journal of Portfolio Management, 15, pp. 4-12.

[5] Dominguez, K. Frankel F. (1993b), " Does Foreign Exchange Intervention Work ? " Institute for International Economics, Washington, D.C.

[6] Dominguez, K. M. (2003), " The market microstructure of central bank intervention ", Journal of International Economics 59(1), pp. 25-45.

[7] Ederington H. Lee H. (1993), " How markets process information : News releases and volatility ", Journal of Finance 48, pp. 1161-1191.

[8] Jones M. Owen L. Robin L. (1998), " Macroeconomic news and bond market volatility ", Journal of Financial Economics, 47, pp. 315-337.

[9] Jorion P. (1996), " Risk and Turnover in the Foreign Exchange Market " dans J. Frankel, G. Galli et A. Giovannini (eds), The Microstructure of Foreign Exchange Markets, Chicago, University of Chicago Press.

[10] Kuttner N. (1999), " Monetary policy surprises and interest rates : Evidence from the Fed funds futures market ", Manuscript, Federal Reserve Bank of New York.

[11] Li L. Engle F. (1998), " Macroeconomic announcements and volatility of treasury futures ", Department of Economics Discussion Paper, University of California, San Diego, pp. 98-27.

[12] Mitchell L. Mulherin J. (1994), " The Impact of Public Information on the Stock Market ", Journal of Finance, 49 (3), pp. 923-950.

[13] Moschitz J. (2004), " Monetary Policy Implementation and Volatility in the Euro Area Money Market ", Money Macro and Finance Research Group (MMF), Conference 2004.

[14] Mussa M. (1981), " The Role of Official Intervention ", Group of Thirty Occasional Papers, 6, Group of Thirty, New York.

[15] Reinhart R. Simin T. (1997), " The market reaction to Federal Reserve policy action from 1989 to 1992 ", Journal of Economics and Business, 49(2), pp. 149-168.

[16] Roley V. Sellon H. (1998), " Market reaction to monetary policy nonannouncements ", Federal Reserve Bank of Kansas City, Research Working Paper, pp. 98-06.

[17] Thornton L. (1998), " Tests of the market's reaction to federal funds rate target changes ", Federal Reserve Bank of St. Louis Review, 80(6), pp. 25-36.

Evaluation des Densités de Prévision: Une Approche GMM

Jaouad Madkour

Laboratoire d'Economie d'Orléans
Rue de Blois - BP 6739
45067 Orléans cedex 2
France

Résumé

La prévision joue un rôle central avant toute prise de décision économique, elle permet en particulier d'orienter les politiques économiques des pouvoirs publics et les projets d'investissement des entreprises. Or, des phénomènes exogènes aléatoires tels que les chocs économiques viennent sans cesse altérer la qualité des prévisions et les rendent incertaines. Par conséquent, il ne suffit pas d'annoncer qu'une variable économique sera établie à un certain niveau sans caractériser en même temps l'incertitude qui règne autour de cette prévision dite "*ponctuelle*".

Il existe deux manières de prendre en compte l'incertitude autour d'une prévision. La première, connue sous l'appellation "*prévision par intervalle*", consiste à construire un intervalle de confiance centré (ou non) sur la prévision ponctuelle. La deuxième, dite "*densité de prévision*", considère toute la distribution de probabilité afin de mieux capter l'effet asymétrique des chocs. Dans ce papier, on s'intéresse à la dernière forme de prévision. Plus précisément, on cherche à évaluer la qualité d'une densité de prévision. Selon Diebold, Gunther et Tay (1998), tester la validité d'une telle prévision revient simplement à tester l'adéquation de la transformée de Rosenblatt des observations hors échantillon, calculée par rapport à cette densité de prévision, à la loi Uniforme sur l'intervalle [0,1]. Les auteurs utilisent un histogramme et un corrélogramme pour vérifier l'uniformité et l'indépendance des variables transformées respectivement. Berkowitz (1999), quant à lui, opère une deuxième transformation en appliquant la fonction quantile de la loi normale centrée et réduite à la transformée de Rosenblatt. Il montre que si la densité de prévision est valide, alors cette nouvelle transformée suit une loi normale centrée et réduite. Il développe un test LR en considérant, sous l'hypothèse alternative, un modèle AR(1) auquel il impose des contraintes pour obtenir les hypothèses nulles de normalité et/ou d'indépendance. Finalement, tester la validité d'une densité de prévision est équivalent à tester une hypothèse distributionnelle. On peut donc exploiter les travaux de Bontemps et Meddahi (2006) qui développent des tests d'adéquation dans un cadre GMM. Les auteurs constatent qu'à certaines lois de probabilité, on peut associer des polynômes orthonormaux. En

particulier, à la loi uniforme sur l'intervalle $[0, 1]$ et à la loi normale centrée et réduite, on associe les polynômes de Legendre et d'Hermite respectivement. De tels polynômes sont d'espérances nulles, on peut donc former des conditions de moment en choisissant certains polynômes et construire des J-statistiques. De plus, sous certaines conditions, la matrice de variance covariance, et subséquemment la matrice des poids optimaux, est connue et est égale à la matrice identité ce qui nous évite une étape d'estimation.

Dans ce contexte, on propose deux tests d'évaluation des densités de prévisions fondés sur la Méthode des Moments Généralisée, l'un pour tester la loi Uniforme sur $[0, 1]$ et l'autre pour tester la loi Normale centrée et réduite. Outre leur simplicité et leur souplesse, des expériences Monte Carlo ont montré que nos tests GMM ont de bonnes propriétés à distance finie comparés au test LR de Berkowitz (1999).

Mots clés: Densité de prévision, GMM, Polynômes de Legendre, Polynômes d'Hermite.

Abstract

Forecast plays a central role before any economic decision-making, it allows in particular to direct the economic policies of public authorities and the projects of investment of companies. Now, unpredictable exogenous phenomena such as the economic shocks ceaselessly come to alter the quality of the forecasts and make them uncertain. Consequently, it is not enough to announce that an economic variable will be established on certain level without characterizing at the same time the uncertainty which reigns around this said "*point forecast*".

There are two ways to take into account the uncertainty around a forecast. The first one, known under the naming "*interval forecast*", consists in building a confidence interval centred (or not) on the point forecast. The second one, said "*density forecast*", consideres the entire probability distribution to get better the asymmetric effect of the shocks. In this paper, we are interested in the last shape of forecast. More exactly, we try to evaluate the accuracy of a density forecast. According to Diebold, Gunther and Tay (1998), testing the validity of such a forecast means simply testing the adequacy of the Rosenblatt transformation of the out of sample observations, calculated with regard to this density forecast, to the Uniform distribution on the interval $[0, 1]$. The authors use a histogram and a correlogram to check the uniformity and the indenpendence of the transformed variables respectively. Berkowitz (1999), as for him, operate a second transformation by applying the gaussian quantile function to the Rosenblatt transform. He shows that if the density forecast is valid, then this new transformed variable follows

a standard Normal distribution. He develops a LR test by considering, under the alternative hypothesis, an AR(1) model in which he imposes some restrictions to get the null hypotheses of normality and/or independence. Finally, testing the validity of a density forecast is equivalent to testing distributional assumption. We can thus exploit the work of Bontemps and Meddahi (2006) developing tests of adequacy in a GMM framework. The authors notice that one can associate orthonormal polynomials to some probability distributions. In particular, to the Uniform on the interval $[0, 1]$ and to the standard Normal ones, one associates the polynomials of Legendre and Hermite respectively. These polynomials have zero expectation, we can thus form moment conditions by choosing some polynomials and construct J-statistics. Furthermore, under some conditions, the variance-covariance matrix, and subsequently the optimal weights one, is known and is equal to the identity matrix. This avoids us a stage of estimation.

In this context, we propose two tests of evaluation of the density forecasts based on the Generalized Method of Moments, the one to test the Uniform distribution on $[0, 1]$ and the other one to test the Normal distribution. Besides their simplicity and their flexibility, Monte Carlo experiments indicate that our GMM-based tests have good finite sample properties when compared to LR test of Berkowitz (1999).

Keywords: Density forecasts, GMM, Legendre polyomials, Hermite polyomials.

Bibliographie

[1] Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 60, 953-966.
[2] Bai, J. (2003), "Testing Parametric Conditional Distributions of Dynamic Models", *The Review of Economics and Statistics*, 85, 531-549.
[3] Berkowitz, J. (1999), "Evaluating the Forecasts of Risk Models", *Federal Reserve Board*.
[4] Berkowitz, J. (2001), "Testing Density Forecasts With Applications to Risk Management", *Journal of Business and Economic Statistics*, 19, 465-474.
[5] Bontemps, C. (2008), "Moment-based Tests for Discrete Distributions", *Working paper*.
[6] Bontemps, C. and Meddahi, N. (2005), "Testing Normality: A GMM Approach", *Journal of Econometrics*, 124, 149-186.
[7] Bontemps, C. and Meddahi, N. (2006), "Testing Distributional Assumptions: A GMM Approach", *Working paper*.
[8] Chatfield, C. (1993), "Calculating Interval Forecasts", *Journal of Business and Economics Statistics*, 11, 121-135.

[9] Chappell, D., Padmore, J., Mistry, P. and Ellis, C. (1996), "A Threshold Model For the French Franc/Deutschmark Exchange Rate", *Journal of Forecasting*, 15, 155164.
[10] Christoffersen, P.F. (1998), "Evaluating Interval Forecasts", *International Economic Review*, 39, 841-862.
[11] Clements, M.P., Franses, P.H., Smith, J. and Dijk, D.V. (2003), "On SETAR Non-Linearity and Forecasting", *Journal of Forecasting*, 22, 359375.
[12] Clements, M.P. and Smith, J. (2001), "Evaluating Forecasts From SETAR Models of Exchange Rates", *Journal of International Money and Finance*, 20, 133148.
[13] Corradi, V. and Swanson, N.R. (2006), "Bootstrap Conditional Distribution Tests In the Presence of Dynamic Misspecification", *Journal of Econometrics*, 133, 779-806.
[14] Crnkovic, C. and Drachman, J. (1996), "Quality Control", *Risk*, 9, 139-143.
[15] Diebold, F.X., Gunther, T.A. and Tay, A.S. (1998), "Evaluating Density Forecasts", *International Economic Review*, 39, 863-883.
[16] Fama, E.F. (1965), "The Behaviour of Stock Market prices", *Journal of Business*, 38, 34-105.
[17] Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054.
Khmaladze, E.V. (1981), "Martingale Approach in the Theory of Goodness-of-fit Tests", *Theory of Probability and its Applications*, 26, 240-257.
[18] Kräger, H. and Kugler, P. (1993), "Nonlinearities in Foreign Exchange Markets: A Different Perspective", *Journal of International Money and Finance*, 12, 195208.
[19] Kupiec, P.H. (1995), "Techniques for Verifying the Accuracy of Risk Measurment Models", *Journal of Derivatives*, 73-84.
[20] Newey W.K. and West, K.D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-708.
[21] Olmo, J. and Escanciano, J.C. (2007a). "Estimation Risk Effects on Backtesting for Parametric Value-at-Risk Models", City University Economics Discussion, Papers 07/11, Department of Economics, City University, London.
[22] Olmo, J. and Escanciano, J.C. (2007b). "Backtesting Parametric Value-at-Risk with Estimation Risk", Caepr Working Papers, Center for Applied Economics and Policy Research, Economics Department, Indiana University Bloomington.
[23] Peel, D. and Speight, A. (1994), "Testing for Non-Linear Dependence in Inter-War Exchange Rates", *Review of World Economics (Weltwirtschaftliches Archiv)*, Springer, 130, 391-417.
[24] Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation", *Annals of Mathematical Statistics*, 23, 470-472.
[25] Wong, C.S. and Li, W.K. (2000), "Testing for Double Threshold Autoregressive Conditional Heteroscedastic Model". *Statistica Sinica*, 10, 173-189.

Testing Interval Forecast: A GMM Approach

Elena-Ivona Dumitrescu & Christophe Hurlin & Jaouad Madkour

Laboratoire d'Economie d'Orléans
Rue de Blois, B.P.6739
45067 Orléans Cedex 2
France

Résumé

La contribution des modèles non linéaires dans la prévision des séries macroéconomiques et financières a été intensivement débattue ces dernières années (voir Teräsvirta, 2006 et Colletaz & Hurlin, 2005 pour une revue de littérature). Comme noté par Teräsvirta, il existe un grand nombre d'études dans lesquelles la performance des modèles non linéaires est comparée à celle des modèles linéaires en utilisant des séries de données réelles. En général, aucun modèle non linéaire (ou linéaire) n'a émergé. Cependant, l'utilisation de modèles non linéaires a conduit au renouveau de la de prévision de court terme, notamment à travers l'émergence de concepts tels que la High Densité Région (Hyndman,1995) ou la densité de prévision par opposition aux prévisions ponctuelles. Par conséquent, ce débat sur la non linéarité et la prévision implique de nouveaux critères de validation des prévisions. C'est le cas des densités de prévision, pour lesquelles des tests spécifiques d'évaluation ont été développés (Bao, Lee et Saltoglu, 2004, Corradi et Swanson 2006 etc).

Par contre, s'il existe de nombreuses méthodes de calcul des HDR et des intervalles de confiance (Chatfield,1993), seules quelques unes proposent des méthodes de validation adaptées à ces types de prévisions. Une principale exception est le papier de Christoffersen (1998) qui introduit des définitions générales d'hypothèses permettant d'évaluer la validité d'un intervalle de confiance obtenu en utilisant n'importe quel modèle (linéaire ou non linéaire). Son approche "*model-free*" est basée sur le concept de violation: on dit qu'une violation s'est produite si la réalisation *ex-post* de la variable n'appartient pas à l'intervalle de confiance *ex-ante*. Trois hypothèses sont donc à distinguer. L'hypothèse de *couverture non conditionnelle* signifie que la fréquence espérée des violations est précisément égale au taux de couverture de la prévision par intervalle. L'hypothèse d'*indépendance* signifie que si l'intervalle de confiance est valide alors les violations doivent être indépendamment distribuées. Autrement dit, il ne doit pas avoir de *clusters* dans la série des violations. Enfin, sous l'hypothèse de *couverture conditionnelle*, le processus des violations satisfait les propriétés d'une différence de martingale. Sur la base de ces définitions, Christoffersen propose un test du Ratio de vraisemblance pour chacune de ces hypothèses en considérant une chaine de Markov binaire sous l'hypothèse alternative. La caractéristique principale

de ce test est que tester la validité d'un intervalle de confiance revient à tester une hypothèse distributionnelle pour le precessus des violations. Si l'on définit une variable indicatrice prenant la valeur un en cas de violation, et zéro dans le cas contraire, il est évident que, sous l'hypothèse nulle de couverture conditionnelle, la somme des variables indicatrices associées à une série d'intervalles de confiance suit une distribution Binomiale.

Dans ce contexte, nous proposons une nouvelle approche GMM pour tester la validité des prévisions par intervalle. En s'appuyant sur le cadre GMM de Bontemps et Meddahi (2005), on définit des J-statistiques basées sur des moments particuliers définis par des polynômes orthonormaux associés à la distribution Binomiale. Une approche similaire a été utilisée par Candelon et al. (2010) dans le contexte du backtesting de la *Value-at-Risk*. Notre approche a de nombreux avantages. Le premier, nous développons un cadre unifié dans lequel les trois hypothèses de couverture non conditionnelle, d'indépendance et couverture conditionnelle sont testées de façon indépendante. Le deuxième, et contrairement aux tests LR, cette approche n'impose aucune contrainte sous l'hypothèse alternative. Le troisième, ce test est facile à mettre en oeuvre et ne pose pas de problèmes de calcul quelle que soit la taille d'échantillon. Le quatrième, cette approche est robuste à l'incertitude des paramètres. Enfin, des simulations Monte Carlo ont montré que, pour des tailles d'échantillon réalistiques, notre test a de bonnes propriétés à distance finie comparé aux tests LR.

Mots clés: Prévision par intervalle, High Density Region, GMM, test d'évaluation, test "*model-free*".

Abstract

The contribution of nonlinear models to forecasting macroeconomic and financial series has been intensively debated (see Teräsvirta, 2006, Colletaz and Hurlin, 2005 for a survey). As noted by Teräsvirta, there is a large number of studies in wich the forecasting performance of nonlinear models is compared with that of linear models using actual series of observed data. In general, no dominant nonlinear (or linear) model has emerged. However, the use of nonlinear models has led to the renewal of the short-terme forecasting approach, especially through the emergence of concepts like High Density Regions (Hyndman, 1995) or density forecasts as opposed to point forecasts. Consequently, this debate on non-linearity and forecasting involves new forecast validation criteria. It is the case of density forecasts, for wich many specific evaluation tests have been developed (Bao, Lee and Saltoglu, 2004, Corradi and Swanson, 2006 etc).

On the contrary, if there are numerous methods to caculate HDR and interval forecasts (Chatfield, 1993), only a few studies propose validation methods adapted to these kind of forecasts. A main exception, is the paper of Christoffersen (1998) which introduces general definitions of hypotheses allowing to assess the validity of an interval forecast obtained by using any type of model (linear or nonlinear). His "*model-free*" approach is based on the concept of violation: a violation is said to occur if the *ex-post* realization of the variable does not lie in the *ex-ante* forecast interval. Three validity hypotheses are then distinguished. The *unconditional coverage* hypothesis means that the expected frequency of violations is precisely equal to the coverage rate of the interval forecast. The *independence* hypothesis means that if the interval forecast is valid then violations must be distributed indenpendently. In other words, there must not be any clusters in the violations sequence. Finally, under the conditional coverage hypothesis, the violation process satisfies the assumptions of a martingale difference. Based on these definitions, Christoffersen proposes a Likelihood Ratio test for each of these hypotheses by considering a binary first-order Markov chain representation under the alternative hypothesis. The main characteristic of this test is that assessing the validity of interval forecasts coms down to testing a distributional assumption for the violation process. If we define a binary indicator variable that takes a value one in case of violation, and zero otherwise, it is obvious that under the null hypothesis of *conditional coverage*, the sum of the indicators associated to a sequence of interval forecasts follows a Binomial distribution.

On these grounds, we propose a new GMM approach to test the interval forecasts validity. Relying on the GMM framework of Bontemps and Meddahi (2005), we define simple J-statistics based on particular moments defined by the orthonormal polynomials associated with the Binomial distribution. A similar approach has been used by Candelon and al. (2010) in the context of the Value-at-Risk backtesting. Our approach has several advantages. First, we develop an unified framework in which the three hypotheses of unconditional coverage, independence and conditional coverage are tested independently. Second, contrary to LR tests, this approach imposes no restrictions under the alternative hypothesis. Third, this GMM-based test is easy to implement and does not generate computational problems regardless of the sample size. Fourth, this approach is proved to be robust to the uncertainty of distributional parameters. Finally, some Monte-Carlo simulations indicate that for realistic sample sizes, our GMM test have good finite sample properties when compared to LR tests.

Keywords: Interval forecasts, High Density Region, GMM, evaluation test, model-free test.

Bibliographie

[1] Berkowitz, J., Christoffersen, F.P. and Pelletier, D. (2010) Evaluating Value-at-Risk Models with Desk-Level Data, *forthcoming in Managment Science*.
[2] Bontemps, C. (2006) Moment-based tests for discrete distributions, *Working Paper TSE*.
[3] Bontemps, C. and Meddahi, N. (2005) Testing normality: a GMM approach, *Journal of Econometrics*, 124, 149-186.
[4] Candelon, B., Colletaz, G., Hurlin, C. and Tokpavi, S. (2010) Backtesting Value-at-Risk: a GMM duration-based test. *forthcoming in Journal of Financial Econometrics*.
[5] Chatfield, C. (1993) Calculating interval forecasts, *Journal of Business and Economic Statistics*, 11, issue 2, 121-135.
[6] Christoffersen, F.P. (1998) Evaluating interval forecasts, *International Economic Review*, 39 , 841-862.
[7] Clements, M.P. and Taylor, N. (2002) Evaluating interval forecasts of high frequency financial data, *Journal of Applied Econometrics*, 18, issue 4, 445-456.
[8] Colletaz, G. and Hurlin, C. (2005) Modèles non linéaires et prévision, Rapport Institut CDC pour la recherche, 106 pages.
[9] Dufour, J-M. (2006) Monte Carlo tests with nuisance parameters: a general approach to finite sample inference and nonstandard asymptotics, *Journal of Econometrics*, 127, issue 2, 443-477.
[10] Hansen, L.P. (1982) Large sample properties of Generalised Method of Moments estimators, *Econometrica*, 50, 1029-1054.
[11] Harvey, D.I. and Leybourne, S.J. (2007) Testing for time series linearity, *Econometrics Journal*, 10, 149-165.
[12] Hyndman, R.J. (1995) Highest-density forecast regions for non-linear and non-normal time-series models, *Journal of Forecasting*, 14, 431-441.
[13] Politis, D.N., Romano, J.P. and Wolf, M. (1999) Subsampling, Springer-Verlag, New-York.
[14] Teräsvirta, T. (2006) Forecasting economic variables with non linear models, in Handbook of Economic Forecasting, G. Elliott, C.W.J. Granger and A. Timmermann editors, Elsevier, volume 1, Chapter 8, 413-457.
[15] Wallis, K.F. (2003) Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting*, 19, 165-175.

Eliciting Individual Preferences for Pension Reform using a Discrete Choice Experiment

Yosr Abid Fourati* and Cathal O'Donoghue**

Abstract

Pension systems have recently been under scrutiny because of the expected population ageing threatening their sustainability. This paper's contribution to the debate is from a political economic perspective as it uses data from a Stated Preferences Choice Experiment to investigate individual preferences for an alternative state pension scheme. Results suggest that individuals' value orientation is an important determinant of their preferences. Respondents' income determines which degree of redistribution is preferred. However, preferences according to age are in contradiction with what is suggested in theory. As far as we know, this is the first attempt in using such a methodology in the field of preferences for redistribution through social security.

Résumé

Les systèmes de retraite ont récemment été scrupuleusement examinés en raison du vieillissement de la population prévu menaçant leur soutenabilité. La contribution de ce papier au débat se fait d'une perspective politico-économique puisqu'il utilise des données d'une enquête de préférences établies du type choix d'options répétées (choice experiment) afin d'investiguer les préférences individuelles pour un plan de retraite public alternatif. Les résultats suggèrent que le système de valeur des individus est un déterminant important de leurs préférences. Le revenu détermine le degré de redistribution préféré. Cependant, les préférences en fonction de l'âge sont en contradiction avec ce qui est suggéré en théorie. A notre connaissance, ceci représente la première tentative d'utilisation d'une telle méthodologie dans le domaine des préférences pour la redistribution à travers les systèmes de retraite.

Mots Clés: Vieillissement de la population, Réforme des systèmes de retraite, Préférences redistributives, Préférences établies, Choix d'options répétées

Keywords: Population ageing, Pension system reform, Redistribution preferences, Stated preferences, Discrete Choice experiments

Acknowledgements: The authors gratefully acknowledge financial assistance from the Department of Economics of the National University of Ireland (NUI), Galway.

*: Corresponding Author, TPAD, Technical and Practical Assistance to Development, 3 rue de la ligue arabe, El Menzah 6, Tunis – Tunisia, yosraf@tpadoffice.com

**Head, Rural Economy and Development Program, Teagasc Mellows Campus, Athenry, Co. Galway, Ireland

1 Introduction

The implementation of pension systems has been one of the most important achievements in terms of social policy in the developed world during the twentieth century. However, discussions on the "pension crisis" and the way to reform pension systems have rapidly emerged due to demographic and budgetary pressures. Pension reform is nowadays in the forefront of the political agenda of many European countries. Still, the reform process is slow, despite the urgency required, partially due to political sensitivity. This is partially and importantly due to the fact that public pension's schemes are redistributive.

The main focus of this paper is to answer the following questions: what are individual preferences over redistribution by state pension systems and what determines these preferences? Several studies within the political economy theory field have attempted to identify the forces that drive the support for income redistribution and the welfare state. Schwarze and Härpfer (2007) summarize these hypotheses into three arguments. The first is an efficiency argument which posits that individuals are willing to pay to reduce the risk associated with their *ex ante* income distribution because they may be risk averse. The second is the self-interest argument which considers that egoistic pecuniary motives are a major determinant of individual preferences. The third argument relates to inequality aversion. The logic, according to which people would support redistributive government policies if they expect to gain from these policies, has been suggested by the median voter model (Meltzer and Richard (1981))[1], as well as by the Esping-Andersen (1985) as the power resource theory. However, both on the theoretical and empirical grounds, the median voter hypothesis has often been questioned (Moene and Wallerstein (2003), Kenworthy and McCall (2008)). Indeed, preferences may include the social status enjoyed by the individual. Individuals may also be inequality averse. Recently, stated preferences techniques have been introduced to analyze preferences for redistribution by the state. For instance, Corneo and Grüner (2002) find empirical evidence that three effects drive support for redistributive policies: the "homo-oeconomicus" effect, the "public values effect" and the "social rivalry effect". On the contrary, Fong (2001) finds little evidence that self interest is an important determinant of demand for redistribution, in his study social preferences are more important.

In the field of distributive preferences for pension programs, it has been shown that social security is supported primarily by self–interested desires on the part of an important proportion of citizens. Old age public pensions create self-interested beneficiaries that might be against the retrenchment of the welfare state (Pierson (1994)). Little empirical studies have conducted to confirm or disconfirm this hypothesis (apart from Lynch (2006)). However, preferences may also be other-regarding, referred to as social preferences. For example, means-tested schemes have proved to be politically sustainable even if they are concentrated on a small range of people. On the empirical side, a strand of the literature has developed recently to analyze the preferences and opinion of citizens concerning pensions. See for instance the Special Eurobarometer survey on Pension Policy and Pension Reform conducted in 2004. Boeri et al. (2001) use stated preferences contingent valuation methods to analyze attitudes towards possible pension reforms in Germany, Italy, France and Spain. van Groezen et al. (2009) analyze the determinants of

[1] Meltzer and Richard (1981) posit that self-interest is a key determinant of attitudes towards redistributive social policies.

people's preferences for particular kinds of pension provision (public, occupational and private) in 15 European countries. Delaney et al. (2006), examine preferences for specific forms of redistribution in Ireland: unemployment payments, old age pensions and child benefit and find support to the self-interested preferences. See also Ferrara (1993); Lynch (2006); van Els et al. (2003); Devroye (2003), Hamil-Luker (2001). In this study we will utilize a preference survey undertaken in Ireland to try to answer these questions.

Old age pensions are central for the Welfare State in Ireland. On the one hand, they fulfil various objectives among which redistribution and poverty alleviation. On the other hand, public pension expenditures represent a large share of social public expenditures. Over the last decade, several reports (especially governmental) and academic research papers aiming at presenting an overview of the Irish pension system and possible alternatives as regard policy options for reform have been published. Recent examples are the two reports prepared under the aegis of the Pensions Board: the "National Pensions Review (2005)" and "Special Savings for Retirement (2006)"; and the "Green Paper on Pensions (2007)". Even though these reports have covered a wide range of issues, people's opinion and choices regarding the future of the pension system have rarely been taken into consideration.

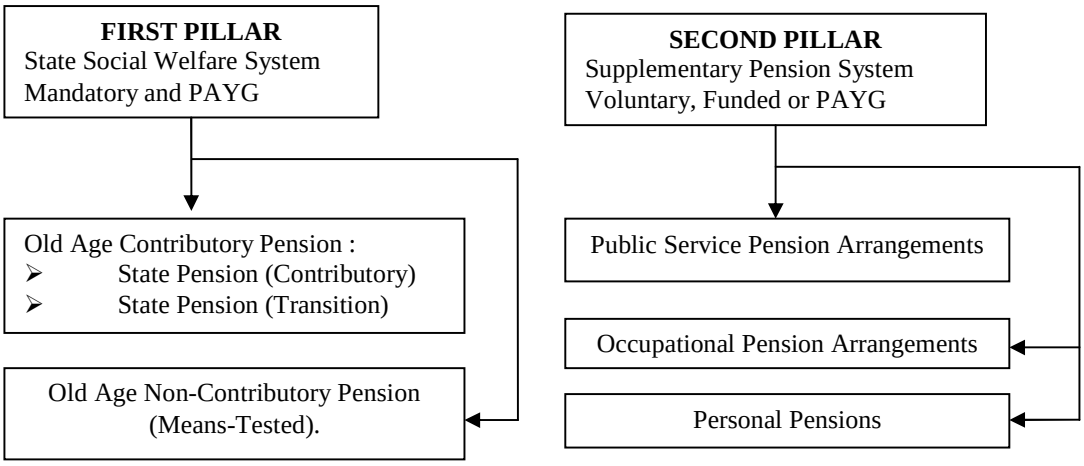
The remainder of this paper is structured as follows. Section two describes the functioning of the Irish pension system, the challenges that threaten its sustainability and the debate around pension reform in Ireland. Section three introduces a life cycle utility model for pension preferences. Section four presents the theoretical basis of discrete choice experiments. Section five outlines the preparation of the questionnaire and presents summary statistics for the sample. In section six, we present results from different discrete choice models. Finally, in section seven, we discuss future directions of research.

2 The challenges of population ageing and poverty for the Irish pension system

2.1 Pension provision in Ireland

The Irish pension system is in many respects typical of the Anglo-Liberal style of welfare state, with a relatively insignificant social insurance system, where means testing and progressive income taxes are more important. The state pension system has its origins in the UK old age assistance act of 1908, with a social insurance pension being introduced for those 70+ in 1961, with a retirement pension at 65 introduced in 1970 (see O'Donoghue (2002). The scheme is Beveridgean in focus with more emphasis on poverty alleviation than on income replacement. Together, the public and private pension schemes operating in Ireland (Figure 1) serve several objectives in different degrees, namely, maintenance of individual standard of living during the period of retirement, poverty alleviation and income redistribution.

Figure 1
Pillars of the Irish pension system



2.2 The old age social welfare system

The public pension system - first pillar, public and mandatory - is funded on a Pay As You Go (PAYG) basis has three components a State pension – Contributory - available from age 66; a State pension – Transition, available from age 65, but requiring retirement; and, a survivor's pension. All are flat rate payments that vary slightly based upon contribution history but independent of earnings, with additional payments for dependants. These are supplemented by means-tested benefits, financed through taxation for those not covered by the insurance system, resulting from significant historical gaps in social insurance coverage. These are nevertheless declining due to the progressive extension of coverage since 1974, resulting in the share of pensioners claiming means-tested payments falling from 45 % in 1994 to 30% in 2004, and expected to reach 14% in 2017. The current replacement rate for a single pensioner is AAA.

2.3 The supplementary pension scheme

The second and third (private and voluntary) pillars are represented by supplementary occupational and/or individual pensions: voluntary tax incentivised private or occupational system, which main objective is to smooth income over an individual's lifetime. Growth in occupational and private pension schemes only accelerated in earnest after the Finance Act 1972 which set up a clear legal and fiscal framework for them, but has levelled out in recent years. After a period of increasing coverage of occupational pensions, ESRI surveys in 1995 found that 52% of the employed workforce was covered by occupational schemes, down from 54.4% 1985, but up from 35.6% in 1974. Recent Statistical Office figures in 2008 indicate that the percentage of employees with an occupation pension has remained relatively constant fallen to about 52%. The proportion of defined contribution schemes has increased from 12% in 1992 (National Pensions Board (1993)) to 17% in 1995 (Hughes and Whelan (1996)) to 33% in 2008 (National

Pensions Board (2008)) of which the majority are in the private sector, where about 50% of the pension members are members of DC schemes.

Personal pension arrangements consist essentially of *Retirement Annuity Contracts* (RACs) used by the self-employed and more recently the *Personal Retirement Saving Accounts* (PRSAs) introduced in 2003 to meet the willingness of the government to extend the supplementary coverage. The main problem related to supplementary arrangements in Ireland concerns the low coverage (around 50 %), especially among self employed. Low contribution rates represent a further inefficiency; they are seen to be insufficient to guarantee adequate replacement rates after retirement. It is therefore not surprising that the government makes a priority of the necessity to increase incentives that would make employers, employees, self-employed and others contribute more into private pension arrangements (both occupational or private). The government is also reviewing its incentives policy to be more effective.

2.4 Poverty among the elderly

Pensioners in Ireland are a vulnerable group due to being highly dependent on transfers payments and having very low labour force participation. In this respect, the *National Anti-Poverty Strategy* (NAPS) adopted by the government in 1997 to address the problem of poverty, and *National Action Plan against Poverty and Social Exclusion* (NAP/incl), identify older people as a particularly vulnerable group to poverty, and both documents list a number of targets in relation to income support and service provision for the elderly. The average net income for a pensioner unit in 2005 was € 327.55 per week (Green Paper (2007)). In 2002, a single pensioner had a replacement rate relative to Gross Average Industrial Earnings of 27.2 % if reliant on means tested benefits and 32 % of Gross Average Industrial Earnings. While coverage of the insurance scheme is increasing, the low replacement rate and lower indexation the level of household earnings growth over the boom years saw a rapid rise in elderly poverty; the relative poverty rate of pensioners rose from 5.9 % in 1994, to 43.3% in 2000 compared with 16.9% for the working age population (DSFA (2002)) and 44.1 % in 2001. Increased indexation of state pensions in the past decade has seen a fall in the poverty rate to 14% in 2006, but rose again in 2007 to 17%.This is high by international standards. In 2005, Ireland had the highest level of relative income poverty in the EU among over-65s (OECD (2005)).

2.5 Population ageing: Key demographic trends

While less serious than some other EU countries, demographic ageing expected to result in increasing the old age dependency ratios, which will put fiscal pressures on the public finances. Nevertheless, the demographic situation is relatively favourable for the next 20 years (Gerald (2004)), especially compared to the situation in the other European countries. Ireland still has a young population, and consequently a longer period to prepare for the transition from low to high dependency. In the OECD countries for instance, the old age dependency ratio is expected to double by 2050, to around 40 percent compared to an average 18 per cent in the 1990s (OECD Social Policy Studies (1996)), whereas in Ireland, it will be rising from 15 percent to 36 percent (Department of Social and Family Affairs (2006))

The increase of the population share of those aged 65 and over represents the main pressure on the public services as age-related public expenditures will have to rise.

Pensions are expected to represent the most important part of these increases. Hence, much of the debate relating to the pensions public policy in Ireland centres around the impact of demographic and economic change on the public finances (DOF (1998)) and the potential cost and funding arrangements. Public spending on first pillar pensions (including public service pensions) is projected to rise from 4.6 % in 2000 to roughly 9-14% of GDP in 2050 (Natali (2004)). Thus the country is faced with both cost and adequacy issues in relation to the pensioner population. Consequently, the combination of the poverty risk among the elderly with the challenging demographic pressures calls for targeted intervention of the State in the field of pensions.

2.6 Reforming the Irish pension system

Because of the concerns raised above about the future increase in public pension expenditures, the National Pensions Reserve Fund Act (2000) established a national pensions fund to help finance both public pensions and public service occupational pensions. Each year, at least 1 per cent of GNP will be deposited in the fund between 2001 and 2055. From 2025 the exchequer will be able to draw down monies from the fund to finance expenditures on public pensions and on the occupational pensions of public sector workers.

There have been a number of structure reforms over the 1990's and 2000's that has resulted in an increase in coverage and since 2001 an increased replacement rate, towards a target of 34 % of average earnings set by the National Pensions Policy Initiative (NPPI[2]). The NPPI also advocated increasing supplementary pension coverage rates with a target coverage rate of 70 % and through increasing personal pension accounts through setting legislative framework to put in place to provide Personal Retirement Saving Accounts (PRSAs); However these reforms have been largely parametric, with policy relying on incrementalism to move towards a universal pension scheme in time rather than a quick move. There seems to be very little public appetite, as manifested in public consultation exercises like the Green paper on pensions in 2007, for major structural reforms such as the move to an earnings related state pension or changes in the state retirement age.

The debate has further been developed more recently[3] focusing on improved adequacy, the abolition of the retirement requirement at age 65 to allow older people to continue to contribute to the economy if they wish to do so, mandatory membership of PRSAs for all workers and a review of the generous tax relief for private pension provision which costs a similar amount as the social welfare pension. The Green Paper on Pensions, 2007 discusses different policy options including the introduction of universal pensions, reforming and back-dating the homemaker's scheme, replacing the average contribution test with a total contribution approach and miscellaneous issues relating to social welfare pensions including indexing, the existence of two contributory pension schemes, social insurance for spouses of farmers/self employed. The Green Paper also considers the introduction of a mandatory or soft-mandatory supplementary pension scheme.

[2] The NPPI has been launched in order to facilitate national debate on how to achieve a developed national policy system and to formulate a strategy and make recommendations for actions needed to achieve the system.
[3] See Submission to the National Pensions Board on the Pensions Review, Combat Poverty Agency, September, 2005.

Despite these reports, reforming Ireland's pension system is still a difficult task. However, it can be made easier by understanding citizens' opinion concerning the size and shape of the welfare state and more generally, by tackling the sources of political conflict over the potential directions of reform and the different approaches to address sustainability. The different alternatives that can be considered range from maintaining the status quo to some option reforms. Note that maintaining the status quo would mean that, in the short to medium term, about 47,000 people on average would remain outside the Social Welfare pensions system (Green Paper on Pensions (2007)). The other options can be divided into enhancing Social Welfare pensions on one hand and encouraging greater personal savings through supplementary pensions on the other hand. Reforming the state pension system implies making it more generous (through improving the adequacy of the system) and less means-tested (through extending coverage). This would require higher public spending on pensions, and thus higher contribution rates. Shifting away from the usual Anglo-Saxon type and implementing earnings-related pension benefits can also be considered. The choice experiment aims at evaluating these reform options from the citizens' point of view.

3 Modelling State Pension System Preferences

How different variables affect people's evaluation of the public pension system? Political economy literature on the determination of pension systems' parameters models individuals' preferences in an overlapping generation setting, where agents choose the values of the parameter(s) through maximizing their utility function over the life cycle. In line with this strand of the literature, we introduce a life-cycle model of pension preferences involving a number of choices:

- the level of the contribution rate,
- the size of the pension benefit,
- how benefits are redistributed,
- the eligibility age for the benefit, and
- the resulting poverty rate among the elderly induced by the pension system chosen.

We consider a two-periods overlapping generation model. Individuals are successively active (18-64 years) then retired (older than 65 years). In addition to age distinction, respondents also differ in revenue endowment. For simplicity we assume that the society consists of two groups of individuals. An individual of type *i* is characterized by his exogenous income level: w_i, $i = m-; m+$ with $m-$ for below the median income, $m+$ for above the median income[4]. Each individual enters working activity at time *0*, retires at date $1-l^o$ and lives until time *1*. We note C_i^g the discounted lifetime income of the respondent of generation g ($g = y, o$) and income group i ($i = m-; m+$). See Table 1 for a definition of the lifetime income of each group.

The lifetime budget constraint of an old agent is given by:

$$C_i^o = c_i^o + (1+\rho)c_i^{y'} = l^o b_i^o + (1+\rho)c_i^{y'} \qquad \textbf{(1)}$$

[4] The annual individual mean income for the sample is €16699 and the annual individual median income is €14000.

The lifetime budget constraint of the young agents is given by:

$$C_i^y = c_i^y + \frac{c_i^{o'}}{1+\rho} = (1-l^o)w_i(1-\tau) + \frac{1}{1+\rho}c_i^{o'} \qquad \textbf{(2)}$$

l^o represents lifetime leisure, it represents the period spend in retirement, ρ is a discount factor, τ is the contribution rate to the state pension scheme, it determines the generosity of the system. In order to concentrate on the redistributive feature of the pension system, we do not consider explicitly the possibility to redistribute income through another tax. c_i^o is consumption during old age and c_i^y is consumption during youth. b_i^o is the level of the state pension benefit for an old individual *i*.

Table 1
Definition of the revenues of the different groups

Pension benefit

	Below median income	Above median income
Young	$\frac{b_{m-}^{o'}}{1+\rho}$	$\frac{b_{m+}^{o'}}{1+\rho}$
Old	b_{m-}^{o}	b_{m+}^{o}

Consumption

	Below median income	Above median income
Young	$c_{m-}^{y} + \frac{c_{m-}^{o'}}{1+\rho}$	$c_{m+}^{y} + \frac{c_{m+}^{o'}}{1+\rho}$
Old	$c_{m-}^{o} + (1+\rho)c_{m-}^{y'}$	$c_{m+}^{o'} + (1+\rho)c_{m-}^{y'}$

A representative individual of generation *g* and ability *i* maximises a utility function $U_i^g(C_i^g)$. Individuals are assumed to be altruistic, that is, they derive utility not only from their own lifetime consumption and leisure but also from the consumption enjoyed by the elderly of the opposite income group, and from preferences regarding the poverty rate among the elderly. That is, a representative individual votes over the pension benefit they would receive from the state pension system, but also on what the current pensioners receive. Indeed, individuals are not purely selfish and they might dislike outcomes that induce high poverty among the elderly, they are poverty averse (or inter-generationally inequality averse). Poverty aversion has an impact on the "size" of the state pension scheme, that is, on "how much is distributed". Furthermore, they have distributional preferences over the public pension system (they might be intra-generationally inequality averse). Attitudes to inequality have impact on "how pensions are distributed", that is, whether pension benefits are means-tested; universal or earnings-related (see Figure 2).

The maximization problem of the old generation is given by:

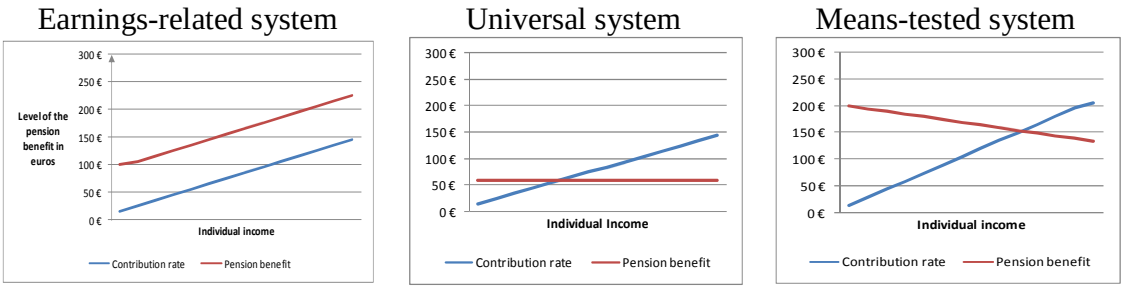
$$\max_{C_i^o} U_i^o = \alpha_i^o C_i^o(b_i^o, \tau) + \beta_i^o(l^o) + \gamma_i^o Pov(c_{m+}^o, c_{m-}^o, \tau) + \theta_i^o u(c_k^o) \qquad \textbf{(3)}$$

Similarly, the maximization problem of the young generation is given by:

$$\max_{C_i^y} U_i^y = \alpha_i^y C_i^y(\tau, b_i^{o'}) + \beta_i^y l^{o'} + \gamma_i^y Pov(c_{m+}^o, c_{m-}^o, \tau) + \theta_i^y u(c_k^o) \qquad \textbf{(4)}$$

$k = m-, m+$, $k \neq i$. α_i^g is the preference parameter associated with lifetime consumption. Pov is the current poverty rate among the elderly. γ_i^g indicates the effect of poverty aversion. Individual feel poverty aversion if $\gamma_i^g < 0$. θ_i^g is the parameter preference for the level of the state pension benefit for the elderly of the other income group. ' refers to the next period variable for the young and the former period variable for the old.

Figure 2
Degree of redistribution of the state pension system



Usually, income is allocated between consumption and saving through a maximisation process of the inter-temporal utility. However, neither data on consumption, nor on saving are available. Therefore, income, during the first period of life is used as a proxy for consumption. We suppose that actual income is equal to disposable income minus taxes (apart from contributions to pensions) and saving. We further assumed that only public pension benefits compose retirement income. This is a restrictive hypothesis as 52.9% of respondents from the sample are members of a private scheme.

The specification above (equations (3) and (4)) implies the estimation of four conditional logit models (see Section 6). The estimation process of the utility model allows testing empirically hypothesis built up from previous literature on demand for redistribution. The question behind the different hypothesis is: what are the forces behind people's preferences for redistribution by the public pension system? The estimation of the life cycle model allows assessing the explanatory power of three different effects. The first hypothesis relates to the income effect, also presented as the "homo economicus effect" in Corneo and Grüner (2002). The first argument stipulates that Income is an important determinant in people's evaluation of the public pension system as respondents are egoistic and will choose the pension alternative that increases their pecuniary gains. The second hypothesis relates to age. Following the assumption made by many political economy models in line with the seminal work of Browning (1975), the age effect suggests that the older the respondent, the more they will choose a higher contribution rate to the public pension system. Similarly, pensioners (who benefit the most from the state pension system) are the more in favour of maintaining the current system and against the "retrenchment" of the welfare state. Hamil-Luker (2001) analyzes the impact of age on public opinion toward government spending on old age assistance and finds that age doesn't have a strong explanatory power. Finally, the value orientation effect suggests that respondents' attitudes towards redistribution through the state pension system are determined by their degree of inequality aversion which stipulates that

individuals are inequality averse, independently from their economic status (see Bowles and Gintes (2000)). This is justified by individuals' altruistic preferences.

Given the hypothesis just mentioned, we are interested in studying how changes in the relevant parameters of the model affect the utility of the different groups in society. Table 2 and Table 3 present some comparative statics properties of the model. The aim is to examine how given utilities change in response to changes in parameters of the variables.

Table 2
Comparative statics of the attributes of the pension system on individual preferences

Utility	Pov	b_{m+}	b_{m-}	τ	l^o
	First derivatives				
U^o_{m-}	-	+	+	-	+
U^o_{m+}	-	+	+	-	+
U^y_{m-}	-	+	+	-	+
U^y_{m+}	-	+	+	-	+

Table 3
Comparison of first derivatives

$\frac{\partial U^o_{m-}}{\partial Pov} < \frac{\partial U^o_{m+}}{\partial Pov}$	$\frac{\partial U^o_{m+}}{\partial b_{m+}} > \frac{\partial U^y_{m+}}{\partial b'_{m+}}$	$\frac{\partial U^o_{m+}}{\partial b_{m-}} < \frac{\partial U^o_{m+}}{\partial b_{m+}}$	$\frac{\partial U^o_{m+}}{\tau} > \frac{\partial U^y_{m+}}{\tau}$	$\frac{\partial U^o_{m+}}{l^o_{m+}} < \frac{\partial U^y_{m+}}{l^o_{m+}}$
$\frac{\partial U^y_{m-}}{\partial Pov} < \frac{\partial U^y_{m+}}{\partial Pov}$	$\frac{\partial U^o_{m-}}{\partial b_{m-}} > \frac{\partial U^y_{m-}}{\partial b_{m-}}$	$\frac{\partial U^o_{m-}}{\partial b_{m-}} > \frac{\partial U^o_{m-}}{\partial b_{m+}}$	$\frac{\partial U^o_{m-}}{\tau} > \frac{\partial U^y_{m-}}{\tau}$	$\frac{\partial U^o_{m-}}{l^o_{m-}} < \frac{\partial U^y_{m-}}{l^o_{m-}}$
		$\frac{\partial U^y_{m+}}{\partial b_{m-}} < \frac{\partial U^y_{m+}}{\partial b'_{m+}}$	$\frac{\partial U^o_{m+}}{\tau} < \frac{\partial U^o_{m-}}{\tau}$	
		$\frac{\partial U^y_{m-}}{\partial b_{m-}} > \frac{\partial U^y_{m-}}{\partial b'_{m+}}$	$\frac{\partial U^y_{m+}}{\tau} < \frac{\partial U^y_{m-}}{\tau}$	

4 Theoretical basis of Discrete Choice Experiments

The demand for reforming the public Irish pension system is assessed through using and extending a particular stated preferences method: choice experiment; conducted to value individuals' preferences regarding the parameters of the Irish public pension system, more precisely through estimating preferences for alternative pension systems (means-tested, universal or earnings-related) as a function of the system attributes and individuals' characteristics and attitudes.

Discrete CEs have widely been used in the literature since their introduction in the marketing and transport fields by Louviere and Hensher (1982) and Louviere and Woodworth (1983). They have now been widely applied in many areas, such as the environmental economics literature (Blamey et al. (1999)), recreation (Hanley et al. (2002)), health (Hall et al. (2004)), transport economics and marketing (See Hensher (1994) and Louviere (1994) for an overview of the use of choice modeling in the field of transportation and marketing respectively). However, the use of stated preferences

techniques in the field of pensions is still scarce. Discrete choice experiments are consistent with the Lancasterian microeconomic approach (Lancaster (1966)) whereby individuals derive utility from the attributes of a good rather than the good itself. In choice experiments, each respondent is typically presented with several choice sets, each containing a set of alternative public goods - in our case alternative public pension systems - defined in terms of their attributes at specified levels. Each alternative being a different combination of the attributes and their levels constructed following a given experimental design Respondents are then asked to choose their most preferred alternative. They are assumed to choose the option that provides them with the highest utility value. A baseline alternative (or status quo) is usually included in each choice set. Typically, five to eight choice sets are included in a choice experiment. The choice experiment reported in this paper contains alternatives that describe hypothetical, constructed potential State pension systems in order to analyze people's preferences for different kinds of public pension provisions, each implying a different degree of redistribution. The hypothetical scenarios were constructed in such a way to be as realistic as possible.

CE finds its theoretical basis in the random utility model (Thurstone (1927); McFadden (1974)), which models choices as a function of attributes and their levels. According to the Random Utility Theory (RUT), the indirect utility function that an individual *k* allocates to one alternative *i* (U_{ki}) is decomposed into two parts: a deterministic part (V_{ki}) which is typically assumed to be linear and additive in the attributes (*x*) of the *A* different alternatives in the choice set, and a stochastic element (ε) which represents the unobservable influences on individual choice.

The indirect utility function associated with alternative *i* for an individual *k* can be written:

$$U_{ki} = V_{ki}(x_{ki}) + \varepsilon_{ki} = \beta x_{ki} + \varepsilon_{ki} \qquad \textbf{(5)}$$

Where β represents the vector of preference parameters (coefficients) associated with the vector of attribute levels x_i

The probability that respondent *k* prefers option *i* to any option *j* in the choice set can be expressed as the probability that the utility associated with alternative *i* exceeds that associated with all other alternatives:

$$P\left[\left(U_{ki} > U_{kj}\right) \forall i \neq j\right] = P\left[\left(V_{ki} - V_{kj}\right) > \left(\varepsilon_{kj} - \varepsilon_{ki}\right)\right] \qquad \textbf{(6)}$$

Assuming that the error terms are independently and identically distributed (IID) with an extreme-value (Weibull) distribution implies that the probability of any particular alternative *i* being chosen as the most preferred can be expressed in terms of the logistic distribution (McFadden (1973)). The following specification is known as a conditional logit model:

$$P\left(U_{ki} > U_{kj}, \forall i \neq j\right) = \frac{\exp\left(\mu V_{ki}\right)}{\sum_{j \in A} \exp\left(\mu V_{kj}\right)} \qquad \textbf{(7)}$$

Where μ is a scale parameter, inversely proportional to the standard deviation of the error distribution. μ is commonly normalized to one (Ben-Akiva and Lerman (1985)). The independence of the Weibull error terms across the different options contained in the choice set implies that specification (3) obeys the Independence from Irrelevant Alternatives (IIA) property (or Luce's Choice Axiom; see Luce (1959)) which states that the relative probabilities of two options being selected are unaffected by the introduction or removal of other alternatives. That is to say that the error terms are uncorrelated between choices and have the same variance (Carson et al. (1994)). Violations of the IIA assumption can be tested using a test developed by Hausman and McFadden (1984). If a violation of the IIA hypothesis is observed, then more complex statistical models, that relax some of the assumptions used, are necessary. Such models are the random parameters logit model (Train (1998)) or the nested logit model (McFadden (1978)). The latter allows for correlations among the error terms within classes of alternatives. The most common method for estimating the parameters of the model is the maximum likelihood.

In order to apply the RUT to our study framework, different attributes have been selected to describe the Irish State pension system, these are described in Table 4.

Table 4
Definition of the choice specific attributes

Attributes	Description
lo_pen	Weekly State Pension for a **low income** individual
hi_pen	Weekly State Pension for a **high income** individual
cost	Weekly contribution to finance State pensions for the **average wage**
pov	Poverty rate among the elderly
ret_age	Minimum **retirement age (years)**

5 Survey Instrument Design and Sample Characteristics

5.1 Questionnaire and sampling design

The survey instrument has been developed following several rounds of testing, involving one pilot test, individual interviews, discussions with PhD students and one focus group composed of participants from the general public. The resulting questionnaire is structured in three parts: Part I: Attitudinal questionnaire; Part II: Choice Experiment; and, Part III: Socio-economic questionnaire. The choice experiment aims at investigating attitudes towards the current Irish state-run pension system and towards some of the likely reforms resulting from the modification of the parameters of the system and involving changes in its principles, degree of redistribution, type of redistribution and poverty among the elderly. As shown in Table 5, five attributes have been selected for valuation. The first attribute is the weekly state pension benefit for a working individual at two third the average earning (low income worker). The second attribute is the weekly state pension benefit for a working individual at three times the average earning (high income worker). The weekly state pension for a low income individual takes on five different levels, among which the first corresponds to the actual state benefit. The levels of the weekly state benefit for a high income individual vary accordingly to describe

either a means-tested, a universal or an earnings-related system. The third attribute is the cost one and corresponds to the weekly contribution amount required to finance state pension benefits for an average individual worker. The fourth attribute is the poverty rate among the elderly implied by each alternative system. Finally, the fifth attribute is the earliest retirement age at which individuals are allowed to receive their State pension, which takes on two values, either the current retirement age: 65 years or a higher retirement age: 68 years. The attributes have been combined in eight choice sets, comprising all the alternatives allowed by the design (32 alternatives). Each choice set comprises four alternative choices. A baseline alternative corresponding to the current system is included in each choice set.

Table 5

Description of the attributes and levels

Attributes	**Levels**				
	Current situation				
Weekly State Pension for a **low income** individual	≤210	210	220	260	180
Weekly State Pension for a **high income** individual	≤210	210, 0, 500	220, 0, 520	260, 0, 620	180, 0, 430
Weekly contribution to finance State pensions for the **average wage**	€30-35	€35-40	€40-45	€45-50	€30-35
		€15-20	€15-20	€20-25	€10-15
		€50-55	€55-60	€65-70	€45-50
Poverty rate among the elderly	Medium (10%-15%)	Low (5%-10%)	Very low (<5%)	Very low (<5%)	Medium (10%-15%)
		High (15%-20%)	Medium (10%-15%)		Very high (20%-25%)
Minimum **retirement age (years)**	65, 68	65, 68	65, 68	65, 68	65, 68

The Sampling strategy was designed so as to obtain a representative sample of the Irish population. The questionnaires were administered using door-to-door technique. In most of the cases, and in accordance with respondents' wishes, the questionnaires were dropped off and collected according to an arranged time. The sampling approach for the survey followed a two-stage procedure. Sampling was first stratified according to the two principle Irish areas classification: Urban/Rural. The second stage involved the choice of locations that are representative of the Irish population in terms of age, gender and socio-economic status. The survey was administered to a representative sample of 498 respondents drawn from the adult Irish population entitled to vote (persons aged 18 years and over). At the end, 326 questionnaires were ready to analyze. The overall response rate of the survey was 65 percent. Table 6 shows the sample age ranges proportions as compared to the national proportions.

Table 6
Proportions of the different age ranges in the total population (sample and national proportions)

	age range						
Area	**18 to 24**	**25 to 34**	**35 to 49**	**50 to 59**	**60 to 69**	**70 or more**	**Total**
Urban							
Number of respondents	54	31	36	32	28	19	200
Percentage in the total sample of the same age range	78.26	62.00	48.65	58.18	60.87	59.38	61.35
Rural							
Number of respondents	15	19	38	23	18	13	126
Percentage in the total sample of the same age range	21.74	38.00	51.35	41.82	39.13	40.63	38.65
Total							
Number of respondents	69	50	74	55	46	32	326
Percentage in the total population sample	21.17	15.34	22.7	16.87	14.11	9.82	100
National proportions in the population aged 15 years and over (%)	18.75	21.40	26.61	14.00	9.63	9.61	100
National proportions in the total population (%)	14.92	17.04	21.18	11.14	7.67	7.65	100
Note: National proportions for the age group 18 to 24 years corespond to the proportions of the age group 15-24.							

5.2 Descriptive statistics for the sample

5.2.1 Profile and main characteristics of the survey respondents

Age is an important variable within the framework of this analysis as pension policy is mainly a "generational" issue. The age distribution of the survey reflects that of the gender population, with around 23% in the 35-49 years age range, which is the largest percentage of any of the age groups. The majority of respondents are male: 52% against and a majority, (47%) attained at least a recognized third level education level. Among respondents, 51% are married. 24% of the individuals live in a household composed of 2 persons, whilst 21% live in a household composed of 4 persons. A majority (27 %) of respondents are private sector employees, while 17% are retired. For 63%, the main household income source is employment, which is the largest proportion of all sources of income, 11% of respondents belong to a household that relies mainly on public pensions for its income and 8% on social welfare. 22% have or have had elderly relatives living in their households.

We expect that pensioners, individuals belonging to a household that derives a large share of his income from public pensions, and individuals who have had elderly relatives living in their households are more likely to oppose retrenchment of public pensions than those who are less dependent on welfare state programs[5]. This hypothesis comes from Pierson's "new politics of the welfare state" which posits that social transfer programs generate self-interested beneficiary groups who will act politically to defend their "programs". This is particularly true for the retirees as they almost completely rely on pensions. Income should have a strong explanatory power in explaining individuals' choices (In accordance with the Meltzer and Richard model). A majority (28.5%) belong

[5] Lynch (2006) tests this hypothesis.

to a household earning between €20000 and €40000 a year before taxes and 22% to a household earning between €40000 and €60000 a year.

5.2.2 Attitudinal profile of the survey sample

People's knowledge about the functioning of the pension system seems to be an age issue implying that the older the respondent, the more they are informed about how the pension system operates and about the level of the pension benefit it provides. While in general, very few people are very well informed (17%), the majority of retirees (+50%) is very well informed. This result is also available for people's knowledge about their (likely) retirement income. We expect better informed voters to be more likely to favour reforms (See Boeri et al. (2001)).

41% of respondents are in favour of a higher pension benefit for the poor (equivalent to a means-tested system) and 37% are in favour of the same pension benefit for everyone (universal system). Only 20% of individuals are in favour of earnings-related pension system. Among these, 51% are respondents who belong to the highest income band, this same income group is the least in favour of the other systems proposed.

Respondents were further presented with three options to deal with demographic ageing. The solution that seems to be the most popular is increasing the retirement age followed by saving more for retirement. The solution that is the least popular is that the government spends more on pensions. Respondents were also asked how they agree with three pension principles. The majority of respondents (36%) strongly disagree with the principle that it is an individual responsibility to save for old age. 47% of respondents strongly agree with the fact that it is the government responsibility to provide each pensioner with a pension benefit. Finally, 41% of respondents slightly agree that the way the pension benefit is provided in Ireland should remain the same, and 31% of respondents strongly disagree with this statement. The general idea from this question is that the status quo (maintaining the current state pension system) is not the preferred option; still a great majority of individuals thinks it is the government responsibility to provide each pensioner with a state pension benefit. Concerning the way to pay for the pension benefit, a large majority of respondents (66%) choose the option: the richer pay proportionally more than the poorer, which suggests that respondents opt for a progressive contribution system.

Besides specific questions on pensions, we also surveyed general attitudes towards the welfare state and demand for redistribution. Concerning income inequality in Ireland, a majority (more than 41%) of respondents strongly agree with the statement "differences in income in Ireland are too large". A clear majority (39.5%) are in favour of less inequality and more tax. A higher proportion of individuals also strongly agree with the fact that it is a governmental responsibility to reduce income inequality.

6 Estimated models and Results

In this section, we report the conditional logit models used to estimate the Random Utility Model (RUM) constructed to assess preferences for alternative state pension systems. The selected choice specific attributes are used to specify the utility of each pension alternative. The models presented below were estimated with STATA using Maximum Likelihood estimation procedures.

6.1 Definition of the Baseline Specifications

We begin by showing four specifications. The four different variants seem satisfactory both on the economic and statistical side. Obviously, the poverty rate among the elderly has a strong effect on people's evaluation of the public pension system. In a first attempt, not reported in the table below, we run a Conditional Logit model, with only regressors the attributes of choices. Results show that the estimate of the cost attribute is positive and the estimates of the attributes "level of the pension benefit for a low income individual" (henceforth low pension) and "level of the pension benefit for a high income individual" (henceforth high pension) are negative. We then estimate a non linear utility function by introducing the square values of cost, poverty, low pension and high pension.

The utility function retained is defined as follows:

$$V_{ki} = \beta_1 lo\_pen_i + \beta_2 lo\_pen_i^2 + \beta_3 hi\_pen_i + \beta_4 hi\_pen_i^2 + \beta_5 \cos t_i + \beta_6 \cos t_i^2 + \beta_7 pov + \beta_8 pov^2 + \beta_9 ret\_age_i \quad \textbf{(8)}$$

Model 1 (CL1) in Table 7 represents the above basic non linear specification. All the coefficients are different from zero at the 1 percent significance level with the exception of the square value of poverty and of the attribute retirement age. The signs of the attributes are as expected. A higher level of the low pension increased the probability that a pension scheme alternative would be chosen, as did the high pension but at a much lower magnitude. Greater poverty and higher cost reduced the choice probability of choosing the associated state pension scheme. A retirement age above the existing one also decreased the probability of the option being chosen. This last effect is however not significant.

In a second model (CL2), we introduce Alternative Specific Constants (ASC's). The number of the ASCs depends on the number of alternatives in a choice set. As we have four alternatives in each choice set, the first alternative being the status-quo, we estimate a set of three ASCs. The ASCs show the effect of asymmetric but unobserved factors of respondents' choices (Morrison et al. (2002)).

The indirect utility associated with choosing alternative *i* is given by:

$$V_{ki} = \beta_1 lo\_pen_i + \beta_2 lo\_pen_i^2 + \beta_3 hi\_pen_i + \beta_4 hi\_pen_i^2 + \beta_5 \cos t_i + \beta_6 \cos t_i^2 + \beta_7 pov + \beta_8 pov^2 + \beta_9 ret\_age_i + \beta_{10} ASC_2 + \beta_{11} ASC_3 + \beta_{12} ASC_4 \quad \textbf{(9)}$$

The parameters associated with the ASCs are positive and significant at the 1 percent significance level. Similarly, all variables (apart from the retirement age) are significant and of the expected signs.

A utility function considering the alternative of choosing the status quo has been estimated (CL3). A new variable, Alternative0, has been created:

$$V_{ki} = \beta_1 lo\_pen_i + \beta_2 lo\_pen_i^2 + \beta_3 hi\_pen_i + \beta_4 hi\_pen_i^2 + \beta_5 \cos t_i + \beta_6 \cos t_i^2 + \beta_7 pov + \beta_8 pov^2 + \beta_9 ret\_age_i + \beta_{10} Alternative0 \quad \textbf{(10)}$$

The Alternative Specific Constant is specified to equal 0 when Alternative A, B or C was selected and to 1 when the "status quo" option was chosen. The alternative specific to the

status quo is negative and significant at the 1 percent significance level, meaning that choosing the current pension scheme decreases utility.

Three variables have been created to account for individuals' choices regarding different kinds of financing state pensions implying different kinds of redistribution: means-tested, universal and earnings-related. The model that has been estimated is given by the following indirect utility function (CL4):

$$V_{ki} = \beta_1 lo\_pen_i + \beta_2 hi\_pen_i + \beta_3 \cos t_i + \beta_4 pov + \beta_5 ret\_age_i + \beta_6 means\_tested + \beta_7 universal + \beta_8 earnings\_related \quad \textbf{(11)}$$

Following CL4, a universal pension scheme increases utility whereas a means-tested or an earnings-related pension scheme reduces the probability of the alternative to be selected. However choosing a universal pension scheme is not significant.

Table 7
Random utility pension choice: Models 1-4

Actual Choice	CL1	CL2	CL3	CL4
State pension benefit	0.136	0.148	0.165	0.013
for a low income individual	(6.38)**	(4.97)**	(7.25)**	(3.12)**
Low pension-squared	-0.0002	-0.0002	-0.0003	
	(-6.08)**	(-4.37)**	(-6.63)**	
State pension benefit	0.012	0.037	0.037	0.003
for a high income individual	(5.71)**	(5.87)**	(5.88)**	(1.11)
High-pension-squared	-0.00001	-0.00003	-0.00003	
	(-7.27)**	(-7.07)**	(-7.08)**	
Poverty rate among the elderly	-11.078	-17.486	-16.775	-4.254
	(-3.88)**	(-4.79)**	(-5.36)**	(-3.28)**
poverty-squared	6.834	8.020	7.535	
	(0.61)	(0.64)	(0.67)	
Cost for an average	-0.149	-0.459	-0.452	-0.039
worker	(-4.32)**	(-5.76)**	(-5.67)**	(-1.14)
Cost-squared	0.0013	0.002	0.002	
	(3.60)**	(5.70)**	(5.59)**	
Retirement age	-0.016	-0.020	-0.022	-0.041
	(-0.94)	(-1.08)	(-1.33)	(-2.39)*
Alternative 2		1.179		
		(4.96)**		
Alternative 3		1.099		
		(4.37)**		
Alternative 4		1.026		
		(4.03)**		
Alternative 1			-0.912	
(current situation)			(-4.18)**	
Means-tested pension alternative				-0.605
				(-1.77)
Universal pension alternative				0.292
				(1.44)
Earnings-related pension alternative				-1.219
				(-2.84)**
log likelihood	-3022.21	-3012.12	-3013.40	-3048.10
Value of z statistics in parentheses; * significant at 5%; **significant at 1%				

All four models predict the expected effects of the pension attributes on the utility of respondents. The levels of the pension benefits (low and high) increase utility associated with the alternative. Cost, retirement age and poverty decrease the probability of the alternative to be chosen. However, what is striking is the strength of the poverty rate in reducing utility. This suggests that individuals are highly poverty averse and are in favour of a system that leads to the lowest poverty rate among the elderly.

6.2 Accounting for respondents' heterogeneity in the choice modelling

In a further step, we propose four groups of models that allow for taking into account respondents' heterogeneity by introducing differences between individuals into the model (Mazzanti (2003)). Specifically, we estimated the impact of age and income on the assessment of the Irish state pension system. Assuming that people are rational utility maximizing, what is referred to as own pecuniary benefits by theoretical literature assume that people should have the strongest preferences for the system that provides them with the highest financial benefit. Age and the degree to which individuals benefit from public pensions are also expected to explain differences in preferences. However, self-interest has been challenged in the empirical literature as not being the main determinant of people's preferences and attitudes. Other forces might increase individuals' utility and thus imply higher demand for redistribution. In this respect, several models have been worked out to include other-regarding motives. For instance, Tabellini (2000) includes altruism from the children to their parents to describe social security systems. Inequality aversion and value orientations may also be important explanatory variables. Van der Heijden et al. (1997) test empirically if altruism and fairness intervene in people's evaluation of public pensions and find strong support that both effects affect individuals' utility.

6.2.1 Testing for Pecuniary Self-Interest

In order to test for heterogeneity in preferences according to income, two separate conditional logit models are run for two separate groups: those reporting an annual individual income above the median annual income of the sample (column 1 in Table 8) and those reporting an annual individual income below the median annual income (column 2 in Table 8). The third column shows the results of interactions between individual income and different types of pension provision.

As in all preceding models, poverty aversion plays a key role in the valuation of the public pension system. As it is expected, a higher poverty rate among the elderly reduces utility of the low-income individuals much more than the utility of the high income individuals, its coefficient is however not significant for the high income group. It is negative and significant at the 1% level for the low income group. A higher low pension benefit increases both the utility of the high income and of the low income group. The coefficients are positive and significant in the two cases; however, the estimated coefficient is higher for the low income group. The estimated coefficients for the high pension are also positive and significant at the 1% level. The estimated coefficient is slightly higher for the high income group. These two effects give strong support to the self pecuniary argument. The estimated coefficients of the cost attribute in both income group models are of the expected sign and significant at the 5% level, however a higher contribution rate seems to decrease the utility of the high income group more than the

utility of the low income group, which is a non expected result. This could be explained by the fact that the richest are not willing to pay more to have a higher pension benefit and probably count more on private pensions. The third column shows that an earnings-related pension system increases individuals' utility as their income increases. Indeed, the interaction between individual income and the earnings-related pension scheme alternative is positive and significant at the 1 percent level. This is a further support for the income effect. Introducing a universal or a means-tested pension scheme impacts utility in the opposite direction: both interactions are negative, with the interaction with the means-tested system being significant at the 1 percent level.

Table 8

Random utility pension choice by income group

Variable	Above median income	Below median income	Interaction with individual income
State pension benefit	0.132	0.141	0.135
for a low income individual	(4.46)**	(4.55)**	(6.15)**
Low pension-squared	-0.0002	-0.0002	-0.0002
	(-4.21)**	(-4.38)**	(-5.93)**
State pension benefit	0.013	0.011	0.009
for a high income individual	(4.52)**	(3.56)**	(3.41)**
High-pension-squared	-0.00001	-0.00001	-0.00001
	(-5.69)**	(-4.59)**	(-5.64)**
Poverty rate among the elderly	-7.543	-15.448	-11.058
	(-1.94)	(-3.65)**	(-3.77)**
poverty-squared	-5.585	22.304	7.807
	(-0.36)	(1.34)	(0.68)
Cost for an average	-0.162	-0.136	-0.134
worker	(-3.44)*	(-2.67)*	(-3.38)**
Cost-squared	0.001	0.001	0.001
	(2.98)*	(2.08)*	(3.49)**
Retirement age	-0.042	0.012	-0.015
	(-1.78)	(0.48)	(-0.88)
Individual income*earnings			0.00001
related pension system			(2.73)**
Individual income*means-tested			-0.00002
pension system			(-4.15)**
Individual income*universal			-0.000003
pension system			(-1.02)
Log likelihood	-1613.24	-1397.56	= -2870.3158
Value of z statistics in parentheses; * significant at 5%; **significant at 1%			

6.2.2 Testing for Age Heterogeneity

Separate conditional logit models have been run for two generations: the young generation (younger than 65 years) and the old generation (over 65 years). For each generation, two models have been run. The first model is the baseline model and the second one includes a variable describing the current pension scheme. Results are reported in Table 9. Almost all pension attributes are statistically significant at the 1% level for all four models, apart from the retirement age which was found to be statistically insignificant in all models.

When including the status quo alternative in the estimation, the variables "low pension" and "cost" become insignificant for the old group, these variables remain however highly significant and of the expected sign for the young group. In all cases, a higher low pension benefit increases the probability of the alternative to be chosen. In fact, estimated coefficients for the low pension are significant at the 1% level for both generations and

higher for the old generation. The older the respondent, the more he/she is in favour of an expansion of the state pension system. Poverty rate among the elderly is negative and significant at the 5% level for the old generation and at the 1% level for the young generation. When not including the current pension system in the utility function, a higher poverty rate decreases the probability of the alternative being chosen at a much higher degree for the old generation than for the young generation. Similarly, in the baseline model, cost decreases the utility of the old generation more than the utility of the young generation. This latter result is in contradiction with the theoretical priors as we would have expected retirees to be in favour of a higher contribution rate as they don't contribute to the system anymore and are net beneficiaries. Finally, the estimated coefficient for the variable "current system" is negative but not significant for the old generation. It is negative and significant at the 1% level for the young generation suggesting that the younger prefer departing from the current state pension scheme. Maintaining the current pension system decreases the utility of the old generation less than the utility of the young generation. However, from theoretical findings, we would expect the current pension system alternative to have a positive effect on the utility of the pensioners and they gain from this system.

Table 9

Random Utility Pension Choice by Generation

Variable	Older than 65 years		Younger than 65 years	
State pension benefit	0.179	0.181	0.129	0.162
for a low income individual	(3.16)**	(3.03)**	(5.58)**	(6.57)**
Low pension-squared	-0.0003	-0.0003	-0.0002	-0.0003
	(-3.06)**	(-3.01)**	(-5.29)**	(-5.88)**
State pension benefit	0.022	0.023	0.010	0.041
for a high income individual	(3.95)**	(1.54)	(4.44)**	(5.78)**
High-pension-squared	-0.00002	-0.00002	-0.00001	-0.00004
	(-4.13)**	(-2.13)*	(-6.08)**	(-6.84)**
Poverty rate among the elderly	-17.181	-17.566	-9.859	-16.755
	(-2.34)*	(-2.20)*	(-3.17)**	(-4.91)**
poverty-squared	17.293	17.226	4.668	5.980
	(0.60)	(0.59)	(0.38)	(0.49)
Cost for an average	-0.295	-0.316	-0.119	-0.484
worker	(-3.34)**	(-1.65)	(-3.18)**	(-5.51)**
Cost-squared	0.002	0.002	0.001	0.002
	(2.32)*	(1.88)	(2.84)**	(5.33)**
Retirement age	-0.008	-0.009	-0.017	-0.025
	(-0.20)	(-0.21)	(-0.93)	(-1.34)
Current pension system		-0.065		-1.091
		(-0.12)		(-4.54)**
Log likelihood	-480.31	-480.31	-2533.87	-2523.46
Value of z statistics in parentheses; * significant at 5%; **significant at 1%				

6.2.3 Testing for Age and Income Heterogeneity in Evaluating the Public Pension System

Table 10 reports the estimation results for four separate groups: old and low income, old and high income, young and low income and young and high income agents.

The estimated coefficients for the high income-old generation are all of the expected sign, but only three variables: the high pension benefit, its squared value and the cost are significant at the 1% level. Poverty rate is significant at the 5% level. The estimated

coefficients for the low income-old generation are also all of the expected sign, apart from the retirement age which is positive. The low pension benefit and its squared value are significant at the 1% level. The high pension benefit and its squared value are significant at the 5% level. All the other estimates are not significant. The estimated coefficients for the high income-young generation are of the expected sign. However poverty rate, its squared value and the retirement age are not significant. Finally, the estimated coefficients for the low income-young generation are all of the expected sign apart from the retirement age which is positive. All the variables are significant apart from the squared value of poverty, the squared value of cost and the retirement age.

The estimated coefficient of the low pension benefit is the lowest for the low income-young generation which is surprising as we expect it to be the lowest for the high income-young generation. However, it is the highest for the low income-old generation, which is as expected. The estimated coefficient for the high pension benefit is also the lowest for the low-income-young group. Poverty aversion is the highest among the high income-old generation, followed by the low income-young generation, suggesting that poverty aversion is still a crucial element in the evaluation of the pension system. Poverty aversion is thus independent from age and income. Cost aversion is also the highest among the high income-old generation. This is also not expected. Indeed, the older should be the more in favour of increasing contribution rates to the public pension system as they are not any more in the tax system; similarly, we expect a higher contribution to decrease the utility of the low income individuals more than the utility of the high income individuals.

Table 10

Estimation Results for Four Groups Differentiated by Age and Income

Variable	over 65 years & above median income	over 65 years & below median income	Below 65 years & above median income	Below 65 years & below median income
State pension benefit for	0.138	0.242	0.132	0.126
a low income individual	(1.85)	(2.73)**	(4.07)**	(3.80)**
Low pension-squared	-0.0002	-0.0005	-0.0002	-0.0002
	(-1.80)	(-2.62)**	(-3.79)**	(-3.67)**
State pension benefit for	0.023	0.019	0.011	0.010
a high income individual	(3.14)**	(2.20)*	(3.44)**	(2.84)**
High-pension-squared	-0.00002	-0.00002	-0.00001	-0.00001
	(-3.05)**	(-2.45)*	(-4.67)**	(-3.86)**
Poverty rate among	-22.741	-6.006	-4.098	-16.667
the elderly	(-2.32)*	(-0.51)	(-0.95)	(-3.65)**
poverty-squared	36.524	-27.450	-16.023	28.869
	(0.97)	(-0.55)	(-0.94)	(1.62)
Cost for an average	-0.322	-0.221	-0.118	-0.118
worker	(-2.80)**	(-1.54)	(-2.25)*	(-2.15)*
Cost-squared	0.002	0.001	0.001	0.001
	(1.83)	(1.05)	(2.12)*	(1.75)
Retirement age	-0.051	0.043	-0.040	0.006
	(-0.87)	(0.66)	(-1.55)	(0.26)
log likelihood	-262.91	-212.40	-1340.08	-1180.32
Value of z statistics in parentheses; * significant at 5%; **significant at 1%				

7 Conclusion

This paper has analyzed the different forces that can affect people's evaluation of the state pension system. Data used to estimate individuals' well-being come from a choice experiment conducted in Ireland in 2008. Respondents were presented with several choice sets, each containing four alternatives from which they had to choose one. Each alternative has different implications for the extent of intra-generational and inter-generational redistribution. In general, within the political economy literature about demand for redistribution though social security, it is assumed that individuals' utility is determined by their self-interest. In this respect, individuals' own characteristics as income and age play a significant role. Nevertheless, other forces, as altruism and social preferences, may also explain people's demand for redistribution. In this respect, estimation results have shown that poverty and inequality aversion indeed affect individuals' utility and their demand for redistribution. All individuals, regardless of age and income are poverty averse and a higher poverty among the elderly decreases utility at an important degree. When the evaluation of the levels of pension benefits depends on age and income of respondents, cost and poverty are independent.

Apart from the two most commonly advanced reasons (age and income) to explain people's heterogeneity in evaluating pensions, people's utility may be affected differently because of value judgments about equality and social justice in general and regarding the objectives and principles of a pension system in particular. Regression results introducing interactions between attitudinal characteristics of the respondents and choice specific attributes for the whole sample suggest that people's attitudes regarding the role of the state in the income redistribution and their opinion about the principles of the pension system are important determinants in their evaluation of the state pension scheme. Several other models not reported in this paper have also been estimated to test for heterogeneity in preferences. Notably, interacting political variables with the choice specific attributes has shown that the level of social contract as well as party partisanship partly explains individuals' heterogeneity in their preferences for different kinds of state pension systems.

One next objective of the study is to derive implicit prices for the pension-specific selected attributes and to investigate individuals' preferences heterogeneity in Willingness to Pay (WTP) for pension reforms. Individual preferences will be aggregated through a majority voting mechanism that handles multidimensional policy issue space (probabilistic voting model) in order to determine pension policy that will be implemented at the national level. Finally, it will be interesting to use the aggregated data from the choice experiment within a micro-simulation framework to account for the impact population ageing on the shape of the public pension system.

8 References

Ageing and Employment Policies: Ireland, (2005), OECD.

Aging in OECD countries: A critical Policy Challenge, (1996), OECD Social Policy Studies, 20.

Ben-Akiva, M., and Lerman, S. (1985), Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Cambridge, MA.

Blamey, R., Gordon, J., and Chapman, R. (1999), Choice Modelling: Assessing the Environmental Values of Water Supply option, The Australian Journal of Agricultural and Resource Economics, 43(3), 337-357.

Boeri, T., Börsch-Supan, A., and Tabellini, G. (2001), Would You Like to Shrink the Welfare State? A Survey of European Citizens, Economic Policy, 16(32), 7–50.

Bowles, S. and Herbert, G. (2000), Reciprocity, Self-Interest, and the Welfare State, Nordic Journal of Political Economy, 26 (1),

Browning, Edgar K. (1975), Why the Social Insurance Budget Is Too Large in a Democracy, Economic Inquiry. 13(3), 373-88.

Carson, Richard T. et al. (1994), Experimental Analysis of Choice, Marketing Letters, 5(4), 351-368.

Corneo, G., and Grüner, H. P. (2002), Individual preferences for political redistribution, Journal of Public Economics, 83, 83-107.

Delaney, L. and O'Toole, F. (2006), Preferences for Specific Social Welfare Expenditures in Ireland, UCD Geary Institute Discussion Paper Series, GEARY WP/8/2006.

Department of Social and Family Affairs, (2002), Sustainable and Adequate Pension Provision for an Ageing Population: Ireland's National Strategy Report to the EU Commission on its Pension System. Brussels: European Commission.

Department of Social and Family Affairs, (2007). Green Paper on Pensions, Dublin.

Devroye, D. (2003), Who Wants to Privatize Social Security? Understanding Why the Poor are Wary of Private Accounts, Public Administration Review, 63(3), 316-328.

DOF, (1998). Long Term Issues Group, Paper. Dublin: Department of Finance.

Eline C.M. van der Heijden", Jan H.M. Nelissenb, andHarrie A.A. Verbona, (1997), Altruism and fairness in a public pension system, Journal of Economic Behavior & Organization, 32, 505-518

Esping-Andersen, G. (1985), Politics against Markets: The Social Democratic Road to Power. Princeton: Princeton University Press.

Ferrera, M. (1993), Citizens and Social Protection: Main Results from a Eurobarometer Survey, Brussels, EC.

Fong, C. (2001), Social Preferences, Self-Interest, and the Demand for Redistribution, Journal of Public Economics, 82, 225–246.

Gerald, J. F. (2004), Ireland – an Ageing Multicultural Economy, The Economic and Social Research Institute, Paper to Merriman Summer School.

Hall, J., Viney, R., Haas, M. and Louviere, J. (2004), Using stated preference discrete choice modeling to evaluate health care programs, Journal of Business Research 57, 1026– 1032.

Hamil-Luker, J. (2001), The prospects of Age War: Inequality between (and within) Age Groups, Social Science Research, 30, 386-400.

Hanley, N., Robert, E. W. and Gary, K. (2002), Modelling Recreation Demand Using Choice Experiments: Climbing in Scotland, Environmental and Resource Economics, 22, 449–466.

Hausman, J. and McFadden, D.L. (1984), Specification tests for the multinomial logit model, Econometrica, 52, 1219-1240.

Hensher, D.A. (1994), Stated preference analysis of travel choices: the state of practice, Transportation, 21, 107-133.

Hughes, G. and Whelan, B. (1996). Occupational and Personal Pension Coverage 1995, Dublin: ESRI.

Kenworthy, L. and McCall, L. (2008), Inequality, public opinion and redistribution, Socio-Economic Review, 6, 35–68.

Lancaster, K. (1966), A new approach to consumer theory, Journal of Political Economy, 74, 132–157.

Louviere, J. and Hensher, D. (1982), On the design and analysis of simulated choice or allocation experiments in travel choice modeling, Transportation Research Record 890, 11–17.

Louviere, J.J. and Woodworth, G. (1983), Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data, Journal of Marketing Research, 20, 350-367.

Louviere, J.J. (1994), Conjoint Analysis. In R. Bagozzi (Ed.) Advances in Marketing Research. Cambridge, MA:Blackwell Publishers.

Luce, R.D. (1959), Individual Choice Behavior: A Theoretical Analysis. New York: John Wiley & Sons.

Lynch, J. (2006), Pension Inequality and Pension Policy Preferences in Europe: Self-Interest, Policy Feedbacks, or None of the Above? Paper prepared for the 16th Conference of the Council of European Studies, Chicago.

Mazzanti, M. (2003) Discrete choice models and valuation experiments. Journal of Economic Studies, 30, 584-604.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, 105–142.

McFadden, D. (1978) Modelling the choice of residential location. Spatial Interaction Theory and Residential Location (Karlquist A. Ed., pp. 75-96). North Holland, Amsterdam.

Meltzer, Allan H. and Richard, Scott F. (1981), A Rational Theory of the Size of Government," Journal of Political Economy, 89, 914-27.

Moene, K. O. and Wallerstein, M. (2003), Earnings Inequality and Welfare Spending, World Politics, 55, 485-516.

Morrison, M., Bennett, J., Blamey, R. and Louviere, J. (2002), Choice Modeling and Tests of Benefit Transfer, American Journal of Agricultural Econonomics, 84(1), 161–170.

Natali, D. (2004), Ireland: The Pension System, European Social Observatory Research Project.

National Pensions Board, (1993), Final Report: Developing the National Pensions System, The Stationary Office. Dublin.

National Pensions Board, (2008), National report and Accounts 2008.

National Pensions Board, (2005), National Pensions Review Report by the Pensions Board to Seamus Brennan TD, Minister for Social and Family Affairs, Dublin.

O'Donoghue, C. (2002), Redistributive Forces of the Irish Tax-Benefit System, Journal of the Statistical and Social Inquiry Society of Ireland, XXXII.

Pierson, P. (1994), Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment, Cambridge: Cambridge University Press.

Schwarze, J. and Häpfer, M. (2007), Are People Inequality Averse, and Do They Prefer Redistribution by the State? Evidence from German Longitudinal Data on Life Satisfaction, The Journal of Socio Economics, 36, 233-249.

Tabellini, G. (2000), A positive theory of social security, Skandinavian Journal of Economics, 102 (3), 523-545.

The Pensions Board, (2006), Special Savings for Retirement. The Pensions Board, Dublin 2

Thurstone, L. (1927), A law of comparative judgment. Psychological Review 34, 273–286.

Train, K.E. (1998), Recreation demand models with taste differences across people, Land Economics 74 (2), 230–235.

Van Els, P.J.A, van den End, W.A. and van Rooij, M.C.J. (2003), Pensions and Public Opinion: a Survey Among Dutch Households, Research Memorandum WO N° 752/Meb-Series N° 2003-18.

van Groezen, B., Kiiver, H. and Unger, B. (2009), Explaining Europeans' preferences for pension provision, European Journal of Political Economy, 25(2), 237-246.

LE CLASSEMENT ET LE REPERAGE SOCIOECONOMIQUE DU MENAGE : CAS DES MENAGES ALGERIENS.

Par Mr. Kahali MAHALI
Enseignant-chercheur, FSEG-UFA (Sétif- Algérie)
« MAHALI.KAMEL@yahoo.fr »

RÉSUMÉ: Un des outils d'observation du social est d'une part la nomenclature des catégories socio professionnelles, si l'étude porte sur l'individu, et d'autre part les catégories socio économiques dans le cas du ménage. Ces deux outils permettent de fournir un découpage de la société selon des critères à la fois économiques et sociaux. Classer les ménages dans des groupes sociaux revient d'un coté à définir les indicateurs et les critères de classification, et d'un autre coté avoir des informations bien précises et fiables sur les éléments à classer. Ces informations concernent les individus constituant le ménage (niveau d'instruction, catégorie socioprofessionnelle, etc.) et le ménage dans ça totalité (nombre de personne par ménage, revenu du ménage, etc.). Donc la construction de la typologie (groupes sociaux) doit passer d'un niveau individuel (Catégories socioprofessionnelles) à un niveau collectif (Catégories socio-économiques). Ainsi le point central de cette étude est de développer la réflexion sur le passage des individus comme entités aux ménages comme autres entités à travers les pratiques algériennes.

MOTS CLES : Catégorisation statistique, catégories socioprofessionnelles, ménages.

Abstract : One of the observation tools of the social environment is the socio-professional categories if the study focuses on the individual and the socio-economic categories in the case of the household. These two tools can provide a segmentation of society according to the criteria of both economic and social. Classify households into groups requires, on one hand, to define the indicators and criteria for classification, and on the other hand, have very specific and reliable information on the elements to be classified. These information concern household members (educational attainment, occupational status, number of persons per household, household income, etc.). So the construction of the typology (social groups) must pass from an individual level (socio-professional categories) at a collective level (socioeconomic categories). Thus, the focus of this study is to develop thinking on the socio-economic classification of household through the Algerian practice.

KEY WORDS: Statistical classification, occupational groups, households.

1. Introduction : Au cours des dernières années notre pays comme tous les pays du monde, a connu des profondes restructurations sociodémographiques et des mutations économiques importantes à tous les niveaux. Il devient désormais primordial d'étudier l'effet de ces transformations socio-économiques sur les algériens prisent individuellement puis collectivement, il s'agit ici de ménages.

Un des outils d'observation du social est d'une part la nomenclature des catégories socio professionnelles, si l'étude porte sur l'individu, et d'autre part les catégories socio-économiques dans le cas du ménage. Ces deux outils permettent de fournir un découpage de la société selon des critères à la fois économiques et sociaux.

2. Cadre conceptuel et contexte de l'étude : Pour repérer les effets des différents bouleversements sur la société, il est nécessaire de déterminer les groupes sociaux des ménages pour lesquels l'observateur (le chercheur) va mesurer les effets. Classer les ménages dans des groupes sociaux revient d'un coté à définir les indicateurs et les critères de classification, et d'un autre coté avoir des informations bien précises et fiables sur les éléments à classer. Ces informations concernent les

individus constituant le ménage (niveau d'instruction, catégorie socioprofessionnelle, etc.) et le ménage dans ça totalité (nombre de personne par ménage, revenu du ménage, etc.). Donc la construction de la typologie (groupes sociaux) doit passer d'un niveau individuel (*Catégories socioprofessionnelles -CSP-*) à un niveau collectif (*Catégories socioéconomiques -CSE-*).

Ainsi le point central de cette étude est de développer la réflexion sur le passage des individus comme entités aux ménages comme autres entités à travers les pratiques algériennes. Les entités ménages, saisies statistiquement par les individus qui les composent, se résument-elles à ceux-ci ? Répondre à cette question a des incidences sur la manière dont a va saisir le ménage : par une somme, par un individu qui le résume. Donc comment saisir les ménages algériens compte tenu de leur structure (rapports matrimoniaux, activité des femmes, activité des autres individus qui composent le ménage comme unité de production économique, des parents, des enfants).

3. Historique de la nomenclature des CSP en Algérie : En Algérie, la nomenclature des catégories socioprofessionnelles a été établie par le Commissariat national au recensement de la population (CNRP actuel ONS) en 1966. Durant les années soixante dix, la construction des CSP a été confiée à l'association algérienne pour la recherche démographique, économique et sociale. La nomenclature produite à été marquée par l'idiologie dominante de cette époque et est toujours en usage à l'Office National des Statistiques à quelques modifications près.

3.1 Principe de construction : La nomenclature adoptée par l'Office National des Statistiques (ONS) a été déterminée à partir la combinaison de trois critères, La situation individuelle, Situation dans la profession et la profession, d'où la logique de catégorisation est la suivante :
1) On distingue les occupés des non occupés. Concernant les non occupés il est distingué entre :
a. Les non concernés c'est-à-dire les enfants de moins de 6 ans non scolarisés ;

b. Les étudiants et écoliers ;

c. Les inactifs qui regroupent les retraités, pensionnés, infirmes, femmes au foyer et les autres inactifs ;

d. Les inoccupés qui regroupent les STR 1 et les STR 2 (sans travail ayant travaillé, n'ayant jamais travaillé).

2) Pour ce qui est des occupés, on tient compte de la situation dans la profession puis de la profession. C'est ainsi que nous distinguons :

Employeurs : Il s'agit des personnes déclarant une situation dans la profession employeur, et ce quelque soit le type d'activité qu'ils exercent. Cette catégorie regroupe les employeurs agricoles et employeurs non agricoles.

Indépendants : Il s'agit de toutes les personnes dont la situation dans la profession est indépendant. Concernant cette catégorie, quatre grands groupes sont à distinguer : les agriculteurs, les commerçants, les artisans et les autres.

Cadres supérieurs et professions libérales : Cette catégorie regroupe, comme son intitulé l'indique, deux types de personnes : les cadres supérieurs et les professions de type libéral (médecins, architectes, avocats, etc.)

Les salariés permanents sont scindés en 11 sous catégories dont les cadres supérieurs, cadres moyens (Techniciens, Enseignement Fondamental, Techniciens de Santé, Administratif et de services), employés (Employés de l'administration, Employés du commerce et des services, Employés des arts et de la culture), ouvriers (Ouvriers non Agricole et Ouvriers Agricoles), et les manœuvres.
Cadres supérieurs : Cette catégorie regroupe deux types profils

a. toutes les personnes justifiant d'un niveau supérieur (licenciés, ingénieurs, docteurs, etc.)

b. toutes les personnes occupant un poste d'encadrement supérieur quelque soit leur niveau (cadres de l'administration, fonctions législatives, cadre supérieur du partie, officiers, etc.)

Cadres moyens : Elle se compose comme suit : Techniciens, Enseignement fondamental, Techniciens de santé, Administratif et de services.

Employés : Cette catégorie contient les employés de l'administration, Employés du commerce & des services et Employés des arts & de la culture.

Manœuvres : Cette catégorie rassemble tous les métiers ne nécessitant aucune qualification (les manœuvres, les dockers et manutentionnaires, les appariteurs, les femmes de ménages, etc.)

Saisonniers : Il s'agit de toutes les personnes dont la situation dans la profession est salarié non permanant et quelque soit leur profession, il s'agit d'une situation dans la profession qui est provisoire.

Apprentis et aides familiaux : Cette catégorie regroupe toutes les personnes déclarant une situation dans la profession apprenti ou aide familial, quelque soit le métier exercé.

4. Classement social des ménages : La nomenclature des catégories socioprofessionnelles permet de classer les individus, elle a aussi l'ambition de définir socialement le ménage. « Mais la question la plus épineuse est celle du classement social des ménages, qu'aucun pays, semble-t-il, n'a su encore convenablement traiter. » (Faucheux, Neyret, 2002, p 134) Nombre des personnes (chercheurs, sociologues …) se sont heurtées à cette question, exprimant une gêne grandissante devant la convention de l'ONS consistant à classer selon la CSP individuelle du chef de ménage.

4.1 Convention de la « personne de référence du ménage 'chef de ménage' » : La CSP permet donc de classer les individus, mais elle peut aussi définir socialement le ménage. « Conçu en effet initialement pour classer les individus selon leur situation professionnelle, le découpage socioprofessionnel a servi rapidement à répartir des ménages, dans la mesure où nombre de pratiques, liées par exemple au revenu ou au logement, résultent du fait que les individus sont presque toujours socialisés dans des unités de base, à l'intérieur desquels ils mettent en commun, au moins partiellement, ressources financières et habitat. » (Desrosières, 1984, p. 8) La catégorie socioprofessionnelle est donc considérée comme une approche empirique de la notion de milieu social. Or le ménage peut être composé de plusieurs individus. La CSP, variable qualitative, ne peut faire l'objet d'une moyenne. Par conséquent, il faut choisir un individu dans le ménage, dont la catégorie socioprofessionnelle représentera le milieu social. Par convention, l'Office National des Statistiques utilise la profession du « chef de ménage », pour déterminer le milieu social du ménage. Le classement d'un ménage selon la catégorie socioprofessionnelle de son chef de ménage «comme personne de référence » pose des problèmes.

a. Des problèmes de définition : Qu'est ce qu'un ménage et qui est son chef? Si l'on s'interroge sur la définition du concept de ménage, il convient de se demander quelles sont les personnes considérées appartenir à un ménage donné : les personnes présentes dans le ménage le jour de l'enquête, ou les personnes vivant habituellement dans le ménage mais pouvant en être absentes pour une durée plus ou moins longue au moment de l'enquête? « La pratique la plus courante consiste à enregistrer chacun dans le ménage où il réside habituellement, mais elle n'est pas totalement généralisée. Dans ce cas le plus fréquent, il se pose alors le problème de définir à partir de critères précis la notion de résidence habituelle. » [1]

Une difficulté voisine tient à la délimitation de la population appartenant à un ménage. Souvent, on parle plus de ménages ordinaires que de ménages, insistant par là sur le fait que les individus résidant dans les communautés (par exemple les prisons) ne font pas partie de la population des ménages. Mais qu'en est-il des étudiants qui résident le plus souvent dans leur établissement

1 Lefranc, C. (1997) *Des difficultés et de l'intérêt de la statistique des ménages*, Statéco, n87-88, août-décembre, p.56

d'enseignement mais peuvent aussi être rattachés à leur ménage d'origine qu'ils fréquentent lors des congés scolaires? Qu'en est-il de même des militaires vivant dans des casernes mais disposant aussi d'une résidence à l'extérieur? Il n'y a pas de façon définitive de trancher ces questions, du moment qu'on prend la précaution d'éviter les doubles comptes lors des traitements statistiques.

Une autre difficulté vient de la distinction entre ménage-logement et ménage-budget, selon la première, il y a équivalence entre les concepts de ménage et de logement, le ménage étant simplement formé de l'ensemble des occupants du logement, utilisé comme résidence principale. La seconde conception du ménage impose en outre aux occupants d'un logement de gérer leur budget en commun pour appartenir au même ménage : selon cette définition, un logement peut donc contenir plusieurs ménages. Dans le contexte algérien, les RGPH font la différence entre les deux définitions contrairement aux enquêtes, la diminution de la taille des ménages et la simplification de leur structure rendant rares les cas où deux groupes de personnes maintenant des budgets séparés partagent le même logement. Même si son usage n'est pas généralisé, il semble que le concept de ménage-logement soit plus facilement utilisable et risque moins de provoquer des erreurs d'interprétation lors de la délimitation des ménages.

Les ménages consistant en des groupes d'individus, les statisticiens trouvent commode d'identifier au sein de chaque ménage un individu servant de point de repère, par rapport auquel peuvent être décrits les liens entre les membres du ménage. Les caractéristiques propres à cet individu, notamment son statut social, doivent aussi servir à situer les ménages les uns par rapport aux autres. C'est pourquoi cet individu doit être une personne "importante" du ménage, soit parce qu'il assure sa subsistance, soit parce qu'il prend les décisions majeures, etc. Pour cette raison, la personne en question est souvent appelée chef de ménage : elle se reconnaît comme tel ou est considérée comme tel par les autres membres du ménage en raison de ses responsabilités. Selon les contextes, la désignation du chef de ménage par l'ensemble des membres du ménage peut aller de soi ou poser plus de difficultés. De même, l'existence de relations assez hiérarchisées au sein des ménages laisse peu d'incertitude sur l'identité du chef de ménage. [2]

Un autre problème provient de la subjectivité liée à la désignation par le ménage lui-même de son individu « chef » : dans des ménages de situations identiques, la désignation d'une personne plutôt que d'une autre peut dépendre de celui qui désigne, ou de considérations internes au ménage (dont le statisticien n'a que faire), même souvent, il y a peu d'ambiguïté quant au choix du chef. [3]

b. Des problèmes liés aux quelques tendances récentes : Le classement d'un ménage selon la catégorie socioprofessionnelle de son chef pose des problèmes liés aux développements de quelques tendances : l'activité féminine dans quelques milieux et le célibat tardif.

Le premier tient à la diversité des structures familiales. Le fait de vivre dans le ménage, d'avoir des enfants diffère selon la catégorie socioprofessionnelle individuelle. Ceci a pour conséquence que les actifs n'ont pas la même probabilité d'être « chef de ménage », déjà en fonction de leur sexe, mais aussi de leur situation professionnelle propre.

Le deuxième problème revient à se poser la question suivante : dans quelle mesure les individus d'un ménage appartiennent-ils au même milieu social ? En effet, l'identité professionnelle est plus ou moins marquée selon les milieux. Dans les catégories où les femmes exercent une activité économique on trouve qu'elles travaillent souvent dans l'enseignement ou dans les services médicaux et sociaux, c'est-à-dire dans des professions du secteur public se définissant par la

2 Surtout dans les pays de l'Afrique, voir (Grenêche, 1995).

3 L'automatisation de désigner le chef de ménage ou la personne référence sera une solution pour gommer l'effet de subjectivité, sachant qu'elle soit bien explicitée.

prédominance du capital culturel. Il est donc clair que l'identification du milieu social d'un ménage à partir de la profession de son chef ne va plus complètement de soi, au moins dans ces catégories.

Le troisième problème conduit à s'interroger sur la pertinence actuelle de la convention adoptée pour les femmes mariées actives. La définition du milieu social du ménage nucléaire par la situation professionnelle de l'homme apparaît aujourd'hui de plus en plus discutable. Il en ressort que pour les couples hétérogames (CSP individuelle différente). La prise en compte de la catégorie socioprofessionnelle de la femme semble nécessaire, au moins en prenant la position sociale la plus élevée dans le couple.

Une tendance récente: Le célibat tardif

Qu'en est-il donc des femmes célibataires ? Qu'en est-il aussi des hommes célibataires tardifs ? Il est ardu de faire admettre qu'un ménage constitué de parents (ou de seulement un des parents) vivant avec leur (ou sa/son) fille ou fils célibataire tardif (ve) peut constituer un ménage nucléaire [4] et l'identifiant par la CSP du père surtout si ce célibataire est actif (ve). Dans le contexte algérien où la famille n'a pas été habituée à vivre le célibat prolongé des hommes, mais plus encore celui des femmes, ce phénomène interpelle. L'identification du milieu social des ménages devrait rendre compte de l'existence de ce phénomène qui gagne en ampleur, au moins pour les ménages où il est plus présent.

3.2 La personne repère du ménage : propositions d'autres définitions.

Il est difficile de classer un ménage sur la base des informations sur tous les membres du ménage. Ainsi, la pratique habituelle est de choisir un membre de ménage comme référence et d'employer l'information sur cette personne pour classer le ménage comme unité. L'idée derrière ceci est que la personne repère est le membre de ménage qui dans un certain sens représente mieux le ménage, par exemple, en exerçant l'influence principale sur la position sociale du ménage et le style de vie, et dont les caractéristiques peuvent être employées pour caractériser le ménage comme unité.

Si nous remettons en cause la définition courante retenue par l'ONS, comment devrait-on la remplacer ? Diverses suggestions peuvent être faites. Les différentes propositions se servent des informations sur le statut d'emploi, le revenu, le sexe et l'âge. Nous proposons aussi de maintenir une certaine combinaison par exemple : le statut et d'âge, la CSP du père et celle de la mère, etc.

a. Quels critères pourraient être employés pour choisir la personne repère ?

Le statut dominant sur le marché du travail, le revenu, le sexe et l'âge sont les critères sur lesquels nos propositions seront formulées. Le revenu et l'âge ont l'avantage d'être des variables continues et pourraient être employés ainsi : personne la plus âgée ou le membre de ménage avec le revenu le plus élevé.

Le critère de revenu : On peut considérer que la personne qui a le plus d'influence sur le ménage est celle qui apporte la plus grande part du budget.
Le critère d'âge : On peut penser que la personne la plus âgée est susceptible également d'avoir une influence importante sur le style de vie et la position sociale du ménage.
Approche de dominance : Cette approche considère la personne repère de ménage comme celle qui est dominante sur le marché du travail, le principe est que la dominance professionnelle reflète une position dominante au sein du ménage.

4 Hadj Ali, D-E. et Lebsari, O. (2006) *La famille algérienne. I. les sources démographiques limites et potentialités*, C.R.E.A.D.

Approche combinatoire : Il semble que les définitions présentées auparavant ont une vision réductrice puisqu'elles réduisent le ménage à la position d'une personne et elles favorisent une forte représentation masculine. De plus, le fait de déterminer le repère uniquement en fonction de contributions financières, ne permet pas de prendre en considération les tâches à l'intérieur du ménage qui ne sont pas d'ordre monétaire et qui sont effectuées principalement par les femmes. Actuellement, on observe au sein des ménages, une certaine division des dépenses.

Une définition qui retient le partage des dépenses conduit à chercher une combinaison pertinente au sein du ménage. Ainsi si on s'en tient au critère relatif aux dépenses et aux situations où plusieurs personnes partagent la charge du ménage. Il serait intéressant d'offrir la possibilité aux ménages de déclarer plus d'un repère, ou encore de n'en nommer qu'un, mais en précisant que cette personne partage ou non les dépenses du ménage avec au moins une autre personne. Cette dernière option aurait l'avantage d'assurer la continuité de la définition actuelle, tout en ajoutant une information pertinente sur le partage des dépenses et des charges. Sur le plan pratique, combiner l'information Père-mère implique tout d'abord de choisir les indicateurs qui seront utilisés (la CSP, la SI, etc.) ensuite de construire une catégorisation combinatoire qui permet de caractériser le ménage à partir des combinaisons possibles. Il s'agit donc de restituer le groupe, d'avoir une approche qui prend en compte les positions des différents membres du ménage.

Enfin, on pense qu'il soit plus pertinent de combiner les CSP au sein d'un ménage.[5] Il serait très important, qu'un ensemble de travaux de recherche tant statistiques que sociologiques puissent être engagés à propos des couples bi actifs et les ménages multi actifs.

5. CONCLUSION : Au terme de cette étude, on a aboutit à une remise en cause de la notion de chef de ménage adopté par l'ONS, pour repérer le ménage, au moins pour les problèmes posées par cette notion. En conclusion, on citera deux recommandations, à notre avis intéressantes applicables pour le cas de l'Algérie, faites par Faucheux et Neyret (1999) dans un rapport intitulé Évaluation de la pertinence des catégories socioprofessionnelles. « Lorsque les comportements sociaux ne sont observables qu'au niveau du ménage dans son ensemble (consommation, épargne, etc.), il importe à tout le moins de croiser la CS de la personne de référence avec le type de ménage (selon le nombre et l'âge des enfants à charge, selon aussi sa position dans le cycle de vie : un couple n'ayant plus d'enfant à charge est à bien distinguer d'un « ménage sans enfant ». Dans la plupart des cas, il semble en outre opportun de reclasser les ménages retraités (y compris les veuves) selon la situation professionnelle qui était celle de la personne de référence quand elle était en activité. »

5. BIBLIOGRAPHIE

[1] Desrosieres, A. et Thevenot, L. (1988) *Les catégories socioprofessionnelles*, Paris : La découverte.

[2] Faucheux, H. & Neyret, G. (1999) *Évaluation de la pertinence des catégories socioprofessionnelles,* Rapport INSEE, D.G.

[3] Hadj ALI, D-E., Lebsari, O. (2006) La *famille algérienne. I. les sources démographiques limites et potentialités,* C.R.E.A.D.

[4] Hammouda, N.E. (2004) *les cadres dans les classifications socioprofessionnelles algériennes (pratique des organismes statistiques),* Cahiers du Gdr Cadres, n°8.

[5] Kieffer, A. (2001) *Catégorisation statistique et harmonisation européenne : l'exemple des CSP*, Correspondances, n°64-65, Tunis, IRMC.

[6] Le numéro de Sociétés Contemporaines consacré aux catégories socioprofessionnelles. (2002) *Sociétés Contemporaines,* n° 45-46.

[7] ONS, (1998) *Code des CSP, code des professions,* Alger.

[5] Au moins pour les ménages nucléaires combiner la CSP du père et celle de la mère si elle est active.

Problème de Bandit Unimodal

Jia Yuan Yu & Shie Mannor

Ecole Normale Supérieure, Département de Mathématiques et Applications,
45 rue d'Ulm, 75005 Paris, France;
HEC Paris, 1 rue de la Libération, 78350 Jouy-en-Josas, France.
jiayuan.yu@ens.fr
&
Department of Electrical Engineering, Technion, Haifa, Israel.
shie@ee.technion.ac.il

Nous considérons le problème de bandit à plusieurs bras [1] dans un cadre où l'ésperance des récompenses est une fonction unimodale par rapport à l'ensemble des bras. En particulier, les bras peuvent appartenir à un intervalle continu ou correspondre aux sommets d'un graphe, où les liens représentent une similarité entre les récompenses. Nous présentons des algorithmes efficaces avec des garanties de performance en rétrospective de la forme de [1, 2]. L'hypothèse d'unimodalité présente un avantage important: nous pouvons déterminer si un bras est optimal en échantillonant les directions autour de celui-ci. Cette propriété nous permet de trouver plus efficacement le bras optimal, ainsi que de détecter plus rapidement des changements abrupts dans les distributions des récompenses. Par example, dans le cas du bandit sur un graphe, l'écart de performance est proportionnel, non au nombre total de sommets, mais au degré maximal et au diamètre du graphe.

Dans un problème de bandit, nous avons un ensemble de distributions $\{P_x : x \in \mathcal{X}\}$ dont les indexes appartiennent à un ensemble $\mathcal{X}$—tel que l'intervalle $[0, 1]$ ou les sommets d'un graphe. Nous choisissons un point X_t à chaque instant et recevons une récompense distribuée selon P_{X_t}. Let but est de trouver l'index avec une récompense espérée la plus élevée possible, tout en prennant le moins d'échantillons possible. L'essence de notre methode pour $\mathcal{X} = [0, 1]$ est l'Algorithme LSE (Algorithm 1), qui emploie la méthode de Kiefer [3] pour trouver le maximum d'un fonction unimodale.

Sous des hypothèses d'unimodalité et de séparation entre les récompenses espérées de bras proches, nous obtenons le résultat suivant.

Théorème 1. *Pour un horizon T fixe, l'Algorithme LSE avec paramètres*

$$\epsilon_n = \frac{C_L \epsilon}{C_H \varphi^3}, \quad \delta_n = \frac{16 \varphi^6 C_H^2}{C_L^2 \epsilon^2 T} \delta \log(8/\delta), \quad \textit{pour tout } n,$$

donne un bras ϵ-optimal avec probabilité $1 - \delta$ après T étapes.

Nous adaptons l'Algorithme LSE aux graphes d'une manière directe en l'appliquant successivement au plus long chemin restant dans le graphe. Nous obtenons alors le résultat suivant.

(Initialisation.) Fixons $[x^L, x^H] = [0, 1]$. Fixons x^M tel que $(x^H - x^M)/(x^M - x^L) = \varphi$, où φ est le nombre d'or. Fixons x^N tel que $(x^N - x^L)/(x^H - x^N) = \varphi$.
Pour $n = 1, 2, \ldots$:

1. Echantillonons les bras $\{x^L, x^M, x^N, x^H\}$ sequentiellement, jusqu'à $(4/\epsilon_n^2)\log(8/\delta_n)$ fois chaque.
2. Dénotons la moyenne empirique des récompenses du bras i par $\hat{r}(i)$. Dénotons le meilleur bras par $x_n^* = \arg\max_{i \in \{x^L, x^M, x^N, x^H\}} \hat{r}(i)$.
3. (Elimination d'intervalle.)
   - Si $x_n^* = x^N$ ou $x_n^* = x^H$, alors éliminons l'intervalle $[x^L, x^M]$. Renouvelons les points $x^L := x^M$, $x^M := x^N$, et $x^N = (x^L + \varphi x^H)/(1 + \varphi)$.
   - Sinon, éliminons $[x^N, x^H]$. Renouvelons les points $x^H := x^N$, $x^N := x^M$, et $x^M = (\varphi x^L + x^H)/(1 + \varphi)$.

Algorithm 1: Algorithme LSE

Théorème 2. *Supposons que la fonction de l'espérance des récompenses est unimodale sur tout chemin du graphe. Soit D_L la plus petite différence entre l'espérance des récompenses de sommets adjacents. Soit d le degré maximal du graphe et ℓ le diamètre du graphe. L'Algorithme LSE Adapté obtient une récompense moyenne espérée d'au moins*

$$\bar{r}(v^*) - \frac{2d\ell}{D_L^2}\log(2(d+1)T) - O(\ell \log T + \ell \log \ell)$$

où $\bar{r}(v^*)$ est la récompense espérée du meilleur sommet.

Mots clés: Problème de bandit, apprentissage séquentiel, fonction unimodale.

Bibliographie

[1] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.
[2] Kleinberg, R. Nearly tight bounds for the continuum-armed bandit problem. *Proceedings of NIPS*, 2004.
[3] Kiefer, J. Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.*, 4(3): 502–506, 1953.

(Initialization.) Set $[x^L, x^H] = [0, 1]$. Set x^M such that $(x^H - x^M)/(x^M - x^L) = \varphi$, where φ is the golden ratio. Set x^N in $[x^M, x^H]$ such that $(x^N - x^L)/(x^H - x^N) = \varphi$.
For iterations $n = 1, 2, \ldots$:

1. Sample the arms $\{x^L, x^M, x^N, x^H\}$ sequentially, until each arm has been sampled $(4/\epsilon_n^2)\log(2m/\delta_n)$ times.
2. Let the sample-average reward of arm i be denoted by $\hat{r}(i)$. Output the arm $x_n^* = \arg\max_{i \in \{x^L, x^M, x^N, x^H\}} \hat{r}(i)$.
3. (Interval elimination.)
   - If $x_n^* = x^N$ or $x_n^* = x^H$, then eliminate the interval $[x^L, x^M]$. Update the points $x^L := x^M$ and $x^M := x^N$, and $x^N = (x^L + \varphi x^H)/(1 + \varphi)$.
   - Else, eliminate $[x^N, x^H]$. Update the points $x^H := x^N$, $x^N := x^M$, and $x^M = (\varphi x^L + x^H)/(1 + \varphi)$.

Algorithm 2: Line search elimination (LSE) algorithm

We consider multiarmed bandit problems where the expected reward is unimodal over the set of arms. In particular, the arms may belong to a continuous interval or correspond to vertices in a graph, where the graph structure represents similarity in rewards. We present efficient algorithms with performance guarantees in retrospect of the type found in [1, 2]. The unimodality assumption has an important advantage: we can determine if a given arm is optimal by sampling the possible directions around it. This property allows us to find more efficiently the optimal arm and detect abrupt changes in the reward distributions. For the case of bandits on graphs, the performance loss is proportional to the maximal degree and the diameter of the graph, instead of the total number of vertices.

In a bandit problem, we are presented with a set of unknown probability distributions $\{P_x : x \in \mathcal{X}\}$ with indices in a set $\mathcal{X}$—such as an interval $[0, 1]$ or the set of vertices of a graph. The learner chooses a point X_t at each time instant and receives a random reward distributed according to P_{X_t}. The goal is to find an index with an approximately highest expected value while taking as few samples from these distributions as possible.

The essence of our method for the case $\mathcal{X} = [0, 1]$ is the LSE Algorithm (Algorithm 2), which employs Kiefer's line search method [3] for unimodal functions.

Under assumptions of unimodal expected reward function and separation of the expected reward of nearby arms, we obtain the following guarantee.

Theorem 1. *For a fixed time horizon T, the LSE Algorithm with parameters*

$$\epsilon_n = \frac{C_L \epsilon}{C_H \varphi^3}, \quad \delta_n = \frac{16 \varphi^6 C_H^2}{C_L^2 \epsilon^2 T} \delta \log(8/\delta), \quad \textit{for all } n,$$

outputs an ϵ-optimal arm with probability $1 - \delta$ after T steps.

We adapt the LSE algorithm to graphs in a straightforward fashion by applying it consecutively to the longest remaining path in the graph. We then obtain the following guarantee.

Theorem 2. *Suppose that the expected reward function be unimodal along every path of the graph. Let D_L be the minimum difference between the expected reward of neighbouring vertices. Let d denote the maximum degree of the graph, and ℓ the diameter of the graph. The Adapted LSE Algorithm achieves an expected average reward of at least*

$$\bar{r}(v^*) - \frac{2d\ell}{D_L^2} \log(2(d+1)T) - O(\ell \log T + \ell \log \ell)$$

where $\bar{r}(v^*)$ is the expected reward of the best vertex.

Keywords: Bandit problems, online learning, unimodality.

EXTENSION DU MODÈLE RLAR(RANDOM LEVEL SHIFT AUTOREGRESSION MODEL): ESTIMATIONBAYESIENNE ET MODÉLISATION DU PRIX DU BARIL DE PÉTROLE.

OUM EL KHEIR MOUSSI

1. ABSTRACT

The application of the bayesian inference is problematic because it is impossible to find a closed form for the posterior distribution in almost cases. This difficulty was overcome due to the development of simulation technics from the 90's, while these methods appeared for the first time in 1974. We state one of these technics which is the Markov Chain Monte Carlo simulation (MCMC).

The basic principle of the MCMC methods is as follow: Starting from an arbitrary value θ_0 we generate a Markov chain $(\theta^t)_t$ which converges in law toward the posterior distribution $p(\theta/\underline{y})$.Producing a Markov chain by an MCMC algorithm is similar to the use of an independently identically distributed sample from $p(\theta/\underline{y})$.Gibbs sampling is a part of MCMC methods. We will explain it (Gibbs sampler) in the case of sampling a model with three parameter dimensions $\theta = (\theta_1, \theta_2, \theta_3)^{'}$.

Let denote by $f_1(\theta_1/\theta_2, \theta_3, \underline{y}); f_2(\theta_2/\theta_1, \theta_3, \underline{y}); f_3(\theta_3/\theta_1, \theta_2, \underline{y})$ the posterior distributions with $\underline{y} = (y_1, y_2, ..., y_n)^{'}$ is a vector of n observations.

The first step of the algorithm consist in choosing a starting point (vector of initial values) for the paramèter θ or generating it in random way $(\theta_{10}, \theta_{20}, \theta_{30})'$.

The second step consist in :$\begin{cases} 1-\text{génerate } \theta_{11}\text{from} f_1(\theta_1/\theta_{20}, \theta_{30}, \underline{y}) \\ 2-\text{génerate } \theta_{21}\text{from } f_2(\theta_2/\theta_{11}, \theta_{30}, \underline{y}) \\ 3-\text{générate } \theta_{31}\text{from} f_3(\theta_3/\theta_{11}, \theta_{21}, \underline{y}) \end{cases}$

Considér the vector (θ_{11}, $\theta_{21}, \theta_{31})^{'}$as initial values vector and return to the step 2.

Générate $(M+N)$ vectors: $(\theta_{11}, \theta_{21}, \theta_{31})^{'}, (\theta_{12}, \theta_{22}, \theta_{32})^{'}, ..., (\theta_{1M+N}, \theta_{2M+N}, \theta_{3M+N})^{'}$

Ignor the first generated M vectors and apply the statistical inference with the last obtained N vectors it means:

$(\theta_{1M+1}, \theta_{2M+1}, \theta_{3M+1})^{'}, (\theta_{1M+2}, \theta_{2M+2}, \theta_{3M+2})^{'}, ..., (\theta_{1M+N}, \theta_{2M+N}, \theta_{3M+N})^{'}$.

Without lost of générality the théory assure that the produced Markov chain is irreducible and converge in law toward the posterior distribution $f((\theta_1, \theta_2, \theta_3/\underline{y})$ when $N \to \infty$.

In this paper we build a derived model from random level-shift autoregressive model (RLAR).

Definition 1. *A set of random variables $(y_t, t \in Z)$ follow a random level shift autoregressive model if it satisfy the following equation:*

$$Y_t = \mu_t + X_t$$

with $\mu_t = \mu_{t-1} + \delta_t\beta_t$ *and* $X_t = \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t$ *where* $(\ \delta_t, t \in Z)$ *is a set of random variables of Bernoulli distribution such that* $P[\delta_t = 1] = p$*,and* $(\beta_t, t \in Z)$ *is a set of observations from a known probability distribution.*

In the previously defined model, the equation can be rewritten as:

$$Y_t = \begin{cases} \mu_{t-1} + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t + \beta_t & \text{with probability } p \\ \mu_{t-1} + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t & \text{with probability } 1-p \end{cases}$$

Remark 1. β_t *can be positif,or négatif. For some applications such the prices,a decrease in the price at a given moment t will be shown by a negatif* β_t *,while an increase will be translated by a positif* β_t*. taking these remarks in consideration we can extent the previous model as follows:*

Definition 2. *A set of random variables* $(y_t, t \in Z)$ *follows the RLAR Bis model if it satisfy the following équation :*

$$Y_t = X_t + Z_t D_t S_t$$

with

-$X_t = \varphi_0 + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t$,

-$(Z_t, t \in Z)$ is a set of Bernoulli distribution such that $P[Z_t = 1] = p$,

-if $Z_t = 1$ then

$$D_t = \begin{cases} 1 & \text{with probability } \psi \\ -1 & \text{with probability } 1-\psi \end{cases}$$

-$(S_t, t \in Z)$ is a set of positives observations, the coëfficients φ_i, i variant from 1 to k satisfy the stationary conditions of stationnarity of an autorégressive process .

-$(\varepsilon_t, t \in Z)$ is a gaussian white noise.

From the previous définition précédente,the process $(y_t, t \in Z)$ satisfait the following équation :

$$Y_t = \begin{cases} \varphi_0 + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t & \text{with probability } 1-p \\ \varphi_0 + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t + S_t & \text{with probability } p\psi \\ \varphi_0 + \sum_{i=1}^{i=k} \varphi_i X_{t-i} + \varepsilon_t - S_t & \text{with probability } p(1-\psi) \end{cases}$$

Under the hypothèsis that $(\varepsilon_t, t \in Z)$ is a gaussian white noise, the distribution of the modèle ca be written as:

$F_{Y_t}(y) = A_1 + A_2 + A_3$ where

$A_1 = (1-p)P[\varepsilon_t < y - \varphi_0 - \sum_{i=1}^{i=k} \varphi_i X_{t-i}]$

$A_2 = p\psi P[\varepsilon_t < y - \varphi_0 - \sum_{i=1}^{i=k} \varphi_i X_{t-i} - S_t]$

$A_3 + p(1-\psi)P[\varepsilon_t < y - \varphi_0 - \sum_{i=1}^{i=k} \varphi_i X_{t-i} + S_t] \Rightarrow$

The density function of the modèle is given by:

$f_{y_t}(y) = B_1 + B_2 + B_3$

$B_1 = \sigma^{-1}(2\pi)^{-1/2}[(1-p)\exp\left(-\frac{(y-\varphi_0-\sum_{i=1}^{i=k}\varphi_i X_{t-i})^2}{2\sigma^2}\right)$

$B_2 = p\psi\exp\left(-\frac{(y-\varphi_0-\sum_{i=1}^{i=k}\varphi_i X_{t-i}-S_t)^2}{2\sigma^2}\right)$

$B_3 = p(1-\psi)\exp\left(-\frac{(y-\varphi_0-\sum_{i=1}^{i=k}\varphi_i X_{t-i}+S_t)^2}{2\sigma^2}\right)$

Then the density function $f_{y_t}(y)$ is a mixture of normal distributions.

For a sample of n observations and conditionally on $x_1 = x_2 = ... = x_k = 0$, the likehood of the modèl will be equal to:

$$f(y_{k+1}, ..., y_n) = \prod_{j=k+1}^{j=n} f(y_j) = \sigma^{-(n-k)}(2\pi)^{-(n-k)/2} \prod_{j=k+1}^{j=n} f_{y_j}(y_j)$$

The paramèters of the model which will be estimated are then:
$p, \psi, \sigma^2, \varphi_0, \varphi_1, ..., \varphi_k, S_{k+1}, ..., S_n$.

In order to estimate theses parameters we use the bayesian approch through the Gibbs Sampling.

For each parameter we derive the posterir distribution in detail . Next we give a practical application using a set of early prices of oil barrel from 1860 to 2007.

The Winbugs package will be used to carry out the simulations.

References

[1]

[2] ALBERT,J.H et CHIB,S.(1993a):Bayes Inférence via Gibbs sampling of autoregressive times series subject to Markov mean and variance shifts. Journal of Business Economic Statistics 11,1-15

[3] CASELLA,G.,and E.I GEORGE.1992:Explaining the Gibbs sampler.The American Statistician,Vol.46,N°3(August 1992)

[4] DEGROOT, M.(1970): Optimal Statistical Decisions ,New York, McGraw

[5] GELLEUX Gilles; Christian ROBERT: Computationnel and inferentiel difficulties with mixture postérior distributions,Journal of American Statistical Association ;Sep.2000, Vol.95, n°451

[6] MOUSSI O:Modélisation de l'évolution des prix du baril de pétrole.Estimation Bayesienne.Thèse de doctorat d'Etat en statistque et économie appliquée (2008)

[7] ROBERT Christian : Bayesian Computationnel methods, CEREMADE Paris Dauphine

[8] ROBERT Christian: Méthodes de Monte Carlo par Chaines de Markov .Ed.Economica

[9] ROBERT Christian,Jean Michel MARIN,Kerrie MENGERSEN: Bayesian Modelling and Inférence on Mixtures distributions,CEREMADE Paris Dauphine

[10] ROBERT Christian,Jean Michel MARIN,:Bayesian Core :A practical approch to computational bayesian statistics,Unversité de Paris Dauphine,2006/2007

[11] SPIEGELHALTER David, ANDREW Thomas, Nicky BEST et Wally GILKS: BUGS 0.5 Bayesian Inférence Using Gibbs Sampler

[12] Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D: WinBUGS – a Bayesian modelling framework : concepts, structure, and extensibility. Statistics and Computing, 10: 325–337.

[13] B. WALSH: Markov Chain Monte Carlo and GIBBS Sampling , avril 2004

[14] LOGICIEL WinBugs: http://www.mrc-bsu.cam.ac.uk/bugs

ENSSEA,11 Chemin Doudou Mokhtar, Ben Aknoun,Alger,Algérie.
E-mail address: eokmoussi@yahoo.fr.

Renforcement de la chaîne logistique par l'entreposage de données trajectoires

Zina Nakhla & Jallel Akaichi
Institut Supérieur de Gestion
41, Avenue de la liberté, cité Bouchoucha, Bardo 2000, Tunisie

Résumé

Pour améliorer la rentabilité et la qualité de services, et faire face à une concurrence de plus en plus ardue, les entreprises sont à la recherche des approches efficaces permettant d'améliorer leurs métiers en générales, et la gestion de la chaîne logistique en particulier qui joue un rôle primordiale dans la réduction des couts, l'amélioration de la qualité des services, et l'augmentation de la productivité. Ce travail s'inscrit dans le cadre d'amélioration de la performance d'une chaîne logistique par la conception d'un Entrepôt de Données Trajectoires (EDT) destiné à collecter les données relatives aux objets mobiles. Les données de l'entrepôt seront analysées pour extraire des connaissances servant à une prise de décision conduisant à renforcer la gestion de la chaîne logistique.

Mots clés : Chaîne logistique, objet mobile, entrepôt de donnée trajectoire

Abstract

To improve the profitability and the quality of services, and to confront the difficult competition, companies are in search of the effective approaches to improve their professions in general, and the management of the supply chain in particular which plays an essential role in reduction of cost, the improvement of the quality of services, and increase of the productivity. This work aims to improve the performance of supply chain by the conception of trajectories data warehouse intended to collect the data relative to the mobile objects. The information stored in the data warehouse will be analyzed to extract knowledge which we use to a decision-making and leading to strengthen the management of the supply chain.

Keyword: Supply chain, Mobile objects, Trajectory data warehouse

1. Introduction

La chaîne logistique et sa gestion, qui est définit par Mentzer [9] comme la coordination systémique, stratégique et la gestion tactique des actions au sein des départements d'une organisation particulière, peut contribuer fortement à la performance de l'entreprise, en permettent de maitriser les dépenses, d'optimiser les flux physiques et d'informations et de limiter les gaspillages. La chaîne logistique est un système caractérisé par la mobilité entre les différents processus. La mobilité dans la chaîne se résume par la fonction de transport. Les véhicules et les chariots présentent les objets mobile, le rôle de transport et de gérer simultanément le flux d'information et le flux physique. C'est est un système qui permette d'apporter une information correcte, où et quand nécessaire, afin de réduire l'incertitude, augmenter la visibilité des produits et augmenter l'efficience globale de la chaîne logistique. L'optimisation des décisions dans la chaîne logistique nécessite le tenir compte des différents éléments qui participent au fonctionnement de la chaîne tel que les produits, les clients, les fournisseurs... Mais les décisions prise ne tiennent pas compte de données relatives aux objets mobiles et sa trajectoire, alors qu'ils peuvent contribuer à l'optimisation de la gestion de la chaîne logistique.

L'objectif de notre travail consiste à améliorer la performance de la chaîne logistique par la conception d'un entrepôt de données trajectoires. Ce dernier sont destinée à collecter les donnée liée aux objets mobile. Ces données seront collecter et analysé pour servir à la renforcement de la gestion de chaîne logistique. Le système décisionnel de l'entreprise basée sur les objets trajectoires, devient de plus en plus agile, adaptable et robuste. Elle est capable de s'adapter rapidement aux changements et aux événements imprévus et imprévisibles, aux opportunités de marché et aux demandes des clients.

Cet article est structuré comme suit: La deuxième section présente un état de l'art sur la chaîne logistique et sa gestion, on présente les principaux travaux de recherche récents traitant la mobilité des chaînes logistique. La troisième section explore les données trajectoires. La quatrième section décrit la conception d'EDT. La cinquième section présente les interrogations. Et finalement une conclusion qui présente les points forts à retenir de notre contribution ainsi que les perspectives liées à ce travail.

2. Etat de l'art

Les travaux sur la mobilité de la chaîne logistique ne tiennent pas compte de la trajectoire des données dans la chaîne logistique dans un contexte décisionnelle. La gestion de la chaîne logistique mobile (mSCM) [4][12] consiste à l'utilisation des dispositifs mobiles pour aider à la conduite des activités de la chaîne logistique et aider les sociétés à réduire les coûts. Le travail de [6] présente un système de coordination de chaîne logistique mobile en temps réel basé sur le principe du système Intelligent Wireless Web (IWW), pour l'amélioration de la communication mobiles et la coordination en temps réel la chaîne logistique. Plusieurs travaux de recherches [10][11] ont présentées le rôle important de la technologie d'identification de fréquence radio (RFID) pour la collecte des données. C'est une technologie qui génère les données relatives aux objets mobiles qui sont utiles pour la prise de décision. L'article [7] développe un système de construction mobile RFID-based de la gestion de chaîne logistique dynamique (le M-ConRDSCM) qui améliore l'acquisition de données en utilisant RFID et fournit aussi un moniteur pour le contrôle. Le travail de Teuteber et Ickerott [12], qui présente une structure et une approche de simulation pour la gestion d'événement de chaîne logistique (SCEM) basé sur auto-ID, les technologies mobiles, les capteurs. Ces technologies peuvent être utilisées ensemble pour prévoir des événements inattendus dans le réseau de chaîne logistique avant qu'ils ne mènent aux problèmes majeurs et fournir une alerte utile. Le système décisionnelle dans la chaine logistique est traité par plusieurs travaux [2][3] en vue d'offrir une aide à la décision et avoir une vue de l'ensemble de l'activité traitée. Le travail de [18], développe un système d'aide à la décision appelé le support de décision pour des chaînes logistiques par la modélisation d'objet ou DEcision Support for Supply Chains through Object Modeling (DESSCOM) qui permet la prise de décisions stratégique, tactique et opérationnelle. Pour l'analyse de ces données collectées, il est nécessaire d'utiliser des outils qui leur permettent de transformer les données brutes dont elles disposent en véritables sources de connaissance. Levray et Mathieu [1] présente les méthodes d'intégration de Data warehouse et le Data mining pour l'optimisation, l'analyse et la prévision dans la chaîne logistique. Ces méthodes se basent sur le principe d'accumulation des données appelée la Technologie de Groupe (GT), qui consiste à regrouper les données et les ressources selon des caractéristiques différentes.

Les travaux de recherches présentées ont traité la mobilité de la chaîne logistique, la collection de données relatives en utilisant les technologies de capteurs (RFID, PDA…). Mais ces travaux n'ont pas prisent en considération l'entreposage et l'analyse de données des objets mobiles dans la chaîne et leur trajectoire. Le concept de la mobilité dans la chaîne logistique a été traité par quelques travaux de points de vue discrets, mais non la trajectoire des objets. Le traitement des informations de la trajectoire est très important, pour la réduction de l'incertitude et l'amélioration l'efficacité de la chaîne par l'analyse de la trajectoire. La solution est de stocker les données de trajectoire dans l'entrepôt de données trajectoire qui est un concept récente relatifs au données trajectoires des objets mobiles.

3. Les objets mobiles

Le développement des technologies mobiles, telles que les téléphones cellulaires et le GPS, machines portables (PDAs) et récemment le RFID (Radio Frequency Identification) ont ouvert la voie vers des applications exploitant la localisation. Il y a une large variété des applications qui manipulent les objets qui changent ses caractéristiques spatiales pendant le temps. Selon [19] un objet mobile est dont la localisation change continuellement dans le temps. Le déplacement d'un objet dans le temps et

l'espace permet d'obtenir une continuité du mouvement. Cette continuité des mouvements [21] présente la trajectoire d'un objet mobile qui est constituée d'un ensemble infini de points, ensemble pour lequel on doit définir une représentation finie. Plusieurs applications se sont intéressées à l'étude de mouvement d'objets mobiles comme l'évolution de trafic [22], la migration d'oiseaux [8]. Et des applications qui ont manipulé les objets mobiles, pour la localisation (tel que la gestion de parcs automobiles, gestion de bateaux…) et d'autre application qui utilise les objets mobiles pour le contrôle et la prévision de tarifs. Ces applications ont généré de nouveaux problèmes qui concernent la gestion d'objets mobiles. La plupart des travaux existants s'inscrivent dans un contexte transactionnel et sont axés sur la modélisation d'objet mobiles et les méthodes d'accès aux données qui concernent ces objets. Comme, il y avait des travaux qui ont concentrées sur les objets mobiles dans le domaine des bases de données [23] et peu des travaux se sont intéressées au domaine d'entrepôt de données [20] où l'exploitation d'historiques d'objets mobiles stockés et les analysées dans un but décisionnel.

4. Modélisation de données trajectoires

Les données de trajectoire incluent par exemple des données des gens en mouvement, des phénomènes naturels comme les tsunamis, aussi bien que des cellules de corps. L'analyse de données de trajectoire permet, par exemple, tirant les modèles comportements des gens. Les modèles peuvent permettre la compréhension de la diffusion de quelques maladies, incitation des mesures appropriées de protéger des populations et empêcher la nouvelle diffusion de la maladie. Dans notre contexte de la chaîne logistique, les données trajectoires incluent les données sur la distance, le durée, la localisation, les arrêts… de trajectoire. L'analyse des trajectoires des moyens de transport dans la chaîne permet d'améliorer les décisions et satisfaire les besoins et les exigences des clients. Cette analyse peut permettre d'assurer la meilleur suivi et le contrôle de processus de transport, prendre des mesures pour éviter tous problèmes prévus tous au long de transportation.

La modélisation de trajectoire dépend de ces types différents, le travail de Spaccapietra [8] distingue trois types de trajectoire (Les trajectoires métaphoriques, Naïve trajectoires géographiques,Les Trajectoires Spatiotemporelles).

Dans notre travail, nous intéressons par le type de trajectoire spatiotemporelle, ce type de trajectoire est utilisé pour montrer la position de l'objet mobile qui est géométriquement représenté comme un point. La trajectoire se compose d'un ensemble de mouvements et un ensemble d'arrêts, pour des raisons différents tel que panne, climat ou l'arrivé à la destination.

En mouvement : Chaque mouvement est caractérisé par un temps de début et le temps de fin. Le début d'un mouvement indique la fin d'un arrêt et la fin d'un mouvement indique le début d'un arrêt, donc l'intervalle de temps d'un mouvement est délimité par deux arrêts consécutifs.

Arrêt : Un arrêt est une partie importante d'une trajectoire où l'objet n'est pas en déplacement. On peut distinguer l'arrêt en trois types d'arrêt :

- Un arrêt planifié : pour la livraison des produits transportés, à l'arrivé à la destination.
- Un arrêt privé : pour avoir une pause, manger …
- Un arrêt imprévu : quand il y a une panne ou un mauvais climat.

Pour analyser la trajectoire de ces objets, il est nécessaire d'enregistrer les données des trajectoires qui se diffèrent selon le statut de l'objet mobile, ainsi que les caractéristiques des objets mobiles.

5. La conception d'EDT pour la chaine logistique

Le développement de nouvelles technologies des dispositifs mobiles engendre le stockage d'un grand volume de données qui concerne les trajectoires des objets mobile. Ce volume de données doit être stocké dans un modèle multidimensionnel pour permettre une analyse précise. Le modèle de l'entrepôt de données classique n'est pas adapté au stockage des informations liées aux objets mobiles. Ces

applications ont besoin d'intégrer des types de donnée spatial et temporel, ce qui nécessite un entrepôt de données Spatiotemporelle, mais cet entrepôt n'est pas suffisant pour le traitement des trajectoires des objets mobile. Il est nécessaire d'utiliser un entrepôt de données qui permet de résoudre le problème de stockage des données des trajectoires. D'où la création de l'entrepôt de données de trajectoire (EDT) qui permet de gérer les caractéristiques de données spatio-temporelles et les dimensions associées. Ce modèle de stockage défini comme un entrepôt de données trajectoire(EDT) dont le but est de stocker les données de trajectoires d'une façon multidimensionnelle. L'analyse de trajectoires permettra aux décideurs d'observer les trajectoires de déplacement des objets mobiles dans la chaîne pour des buts de processus décisionnel. La modélisation de notre EDT nécessite la précision de contenu de l'entrepôt selon les besoins de la chaîne logistique et organiser selon les résultats attendus et les statues de l'objet mobile. On a utilisé pour la modélisation multidimensionnelle de l'EDT le schéma en étoile. Notre EDT pourrait avoir les dimensions suivantes : localisation (dimension spatiale), temps, destination, sous trajectoire, arrêt, mouvement, objet mobile.

Objet mobile : Les objets mobiles assurent le déplacement des marchandises entre les différents services de la chaîne, ces objets permettent aussi de transférer les données nécessaires pour la prise de décision. Le type d'objet mobile peut être soit un véhicule pour le transport externe, soit un chariot pour le transport interne.

Trajectoire : La trajectoire est le chemin parcouru par un objet mobile. La fin de la trajectoire présente l'arrivé à la destination. Trajectoire= ∑ Sous trajectoire

Sous-trajectoire : Chaque sous-trajectoire se compose d'un ensemble de mouvement. Un sous trajectoire limité par deux arrêts successifs. Sous trajectoire= ∑ mouvement

Mouvement : Chaque mouvement d'objet mobile dans une trajectoire a un identifiant et caractérisé par le temps et la localisation de mouvement.

Arrêt : Les arrêts dans une trajectoire sont identifiés et caractérisé par le temps et la localisation de l'arrêt. Un arrêt peut être pour plusieurs raisons soit pour de raison de panne, arrivé à sa destination...

Destination : La destination de la trajectoire peut être soit les consommateurs, les fournisseurs et les processus de la chaîne, soit les services d'un processus de la chaîne.

Nature : La dimension nature précise le caractéristique de la nature de chaque sous trajectoire (urbain, montagne, désert…), cette classe regroupe la nature des endroits visités.

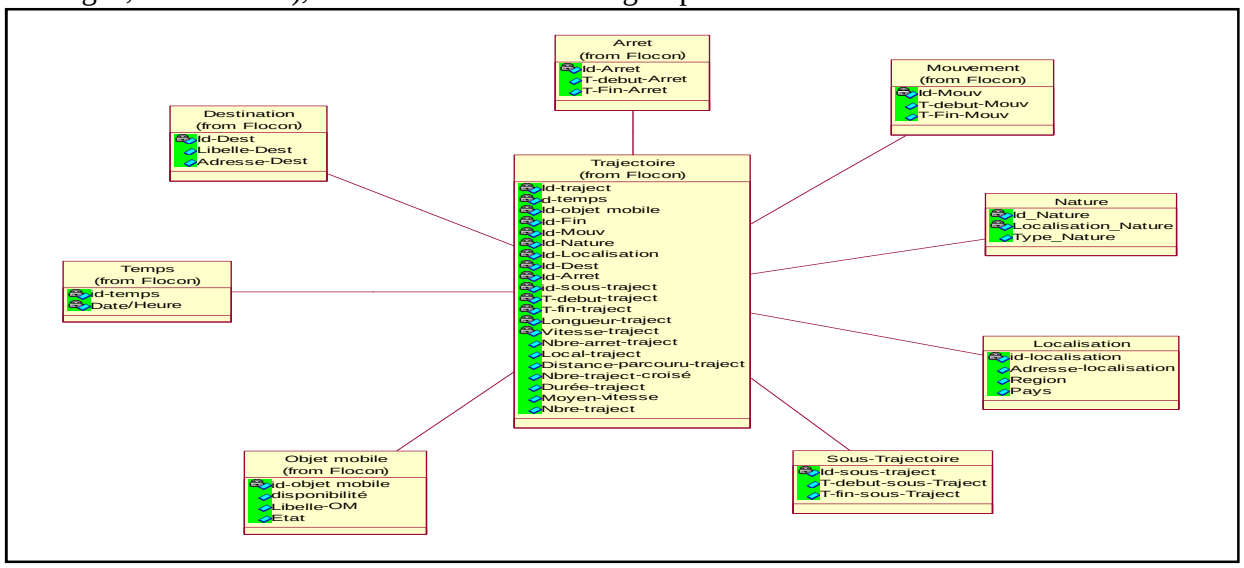


Figure 2 : Le schéma en étoile

6. L'interrogation de l'EDT

Dans cette section nous analysons les données trajectoires relatives aux objets mobiles selon des dimensions spatiales (par exemple la localisation de l'objet mobile) et des dimensions spatiotemporelles (par exemple le changement de la position d'objet mobile en temps et en espace). De plus, l'analyse de trajectoire doit respecter les conditions comme : le destinataire, les caractéristique des trajectoires, le climat…

Plusieurs travaux de recherches ont traités le problème d'interrogation les objets mobiles dans les entrepôts de données. En fait, le travail de Wolfson [13] propose un modèle pour préciser la position incertaine d'un objet mobile en utilisant les requêtes d'interrogations basé sur des formules mathématiques, ces requêtes permettent de préciser la position courante et la position future d'un objet mobile. Le travail de [14] propose un langage d'interrogation pour l'Objet Mobile Spatiaux Temporelles (MOST) appelé le modèle Logique Temporelle Future (FTL). Les auteurs distinguent trois types de questions : instantané, continu et persistant. L'article de Mouza et Rigaux [15] propose un langage d'interrogation basé sur des expressions régulières, destinées à l'obtention d'un modèle appelé « mobility patterns ». Cependant, cette langage est seulement adapté pour les données de trajectoire et ne relie pas la trajectoire avec son environnement, ainsi les classes d'interrogation de ce modèle sont limitées. Le travail de Guting [16], présente un langage d'interrogation pour les objets mobiles dans des réseaux. Ils ont défini deux types de données qui sont gpoint et gline pour représenter le réseau et ils ont défini des types d'opération. L'interrogation de données relatives aux objets mobiles de la chaîne logistique stockée dans l'EDT, est une étape importante pour déduire les connaissances nécessaires à la prise de décision. Ces interrogations sont classées sous deux catégories :

- Interrogation pour l'objet mobile : ces interrogations servent à préciser les données relatives aux objets mobiles (la localisation de l'objet mobile à un instant bien déterminée, les détails concernant l'utilisation des objets mobile dans une période bien déterminé, les objets mobiles disponibles à l'utilisation, les endroits visités par l'objet...). Ces données améliorent la gestion des objets mobiles dans les entreprises.
- Interrogation pour la trajectoire des objets mobiles : ces interrogations concerne les trajectoires des objets mobiles et ces caractéristiques (le temps d'arrêt d'une trajectoire, la durée, la longueur de trajectoire, les trajectoires croisés…). Ces données améliorent la gestion des mouvements des objets mobiles, ainsi que le fonctionnement des moyens de transport.

Ces deux catégories d'interrogations offrent aux décideurs une vue globale sur l'aspect mobile dans une chaine logistique. Les données extraites à partir de ces interrogations servent à renforcer la chaine logistique par l'amélioration des décisions prises qui tiendra compte des données trajectoires.

Le schéma de l'entrepôt de données et les interrogations proposés donnent aux décideurs une vue globale sur l'aspect mobile dans la chaîne en précisant le comportement de l'objet mobile et les caractéristique de sa trajectoire tel que le nombre des arrêts dans une trajectoire, la localisation des trajectoires, la distance parcouru et la durée des trajectoires…

7. Conclusion

Le développement des techniques de suivi de localisation des objets mobiles et la collection de donnée relatifs au mouvement de ces objets mène à une abondance de données relatifs aux objets mobiles et soulève la question de les exploiter pour l'aide à la décision. En effet, l'historique des objets mobiles peut livrer des informations précieuses si on lui applique une analyse multidimensionnelle à différentes échelles.

Dans notre travail on traite le cas de la chaîne logistique, qui est un système caractérisé par la mobilité entre les différents processus de la chaîne ainsi qu'au sein des processus. Notre travail étudie le problème de l'entreposage et l'analyse de données relatives aux objets mobiles dans une chaîne logistique. Les données de ses objets sont stockés dans entrepôt de données trajectoire qui supporte le

volume important des informations liées aux trajectoires des objets mobiles. L'objectif de notre travail consiste à améliorer la performance de la chaîne logistique. On propose un modèle conceptuel pour l'entrepôt de donnée trajectoire qui supporte les données trajectoire dans une chaîne logistique. Les données de l'entrepôt sont analysées pour extraire des connaissances qui servent comme un support à la prise de décision conduisant à renforcer la gestion de la chaîne logistique.
Nous n'avons pas pris en compte dans les interrogations sur l'information future comme l'emplacement futur de l'objet mobile, donc ce type de question peut être traité dans des travaux futurs.

Bibliographie

[1] Levray, R., Mathieu, R. (2009). 'Data warehousing and mining in supply chain'. Encyclopedia of Data Warehousing and Mining p 586-591.

[2] Vural, E., Sengül, O., Davis, S., Günther, H.(2006). 'Business Intelligence for supply chain management system'. Issues in Information Systems, Volume VII, No. 2.

[3] Stefanovica, N., Radenkovicb, B., Stefanovicc, D. (2006).' Supply chain intelligence'. Information Systems Division.

[4] Eng, T.Y.(2006). 'Mobile supply chain management: Challenges for implementation'. Technovation 26 (2006) 682–686.

[6] Soroor, J., Tarokh, M. J., Shemshadi, A.(2009).' Initiating a state of the art system for real-time supply chain coordination'. European Journal of Operational Research 196 (2009) 635–650.

[7] Wang, L.C., Lin, Y.C., Lin, P.H.(2007).'Dynamic mobile RFID-based supply chain control and management system in construction'. Advanced Engineering Informatics 21 (2007) 377–390.

[8] Spaccapietra, S., Parent, C., Damiani, M.L., de Macedo, J.A., Porto, F., Vangenot, C.(2007).'A Conceptual View on Trajectories'. Technical Report, Ecole Polythecnique Federal de Lausanne, 2007.

[9] Mentzer, J.,T. (2001).'Supply Chain Management'.Library of congress cataloging inpublcation data

[10] Saygin, C., Sarangapani, J,. Grasmann, S.E.(2007).'A système approach to viable RFID implementation in the supply chain'. Trends supply chain design and management p 3-28.

[11] Gaukler, G.M., Seifert, R.W.(2007).'Applications of RFID in supply chain'. Trends supply chain design and management p 29-48.

[12] Teuteberg, F., Ickerott, I. (2007). 'Mobile supply chain events management using Auto-Id and sensor technologies'. Trends supply chain design and management p 93-125.

[13] Wolfson O., Chamberlain S., Dao S., Jiang L., Mendez G. (1998). 'Cost and Imprecision in Modeling the Position of Moving Objects'. In: Proceedings of ICDE 1998.

[14] Sistla, P., Wolfson, O, Chamberlain, S., Dao, S.(1997). 'Modeling and Querying Moving Objects'.

[15] Mouza, C., and Rigaux, P.(2005). 'Mobility patterns'. Geoinformatica, 9(23):297–319, 2005.

[16] Guting, R.H., Almeida, V.T., Ding, Z.(2004). ' Modeling and Querying Moving Objects in network'. Technical Report 308, Fernuniversitat Hagen, Fachbereich Informatik.

[18] Biswas, S., Narahari, Y.(2002). 'Object oriented modeling and decision support for supply chains'. European Journal of Operational Research.

[19]Wan, T., Zeitouni, K.(2006). 'Vers un entrepôt d'objets mobiles contraints par le réseau'.

[20]Wan, T., Zeitouni, K. (2005). 'Modélisation d'objet mobile dans un entrepôt de données'. Edition Cépaduès.

[21]Mouza, C., and Rigaux, P.(2005). 'Mobility patterns'. Geoinformatica, 9(23):297–319, 2005.

[22] Brakatsoulas, S., Pfoser, D., Tryfona, N. (2004). 'Modeling, Storing and Mining Moving Object Databases'. Proceedings of the International Database Engineering and Applications Symposium.

[23] Meng, X., Ding, Z.(2003). 'DSTTMOD: A Discrete Spatio-Temporal Trajectory Based Moving Object Databases System'. LNCS 2736.

Inférence asymptotique dans les processus ARCH(q) périodiques

Ines Lescheb

Département de mathématiques, Université Mentouri, 25000 Constantine, Algeria.
e-mail: i.lescheb@gmail.com

Résumé

On étudie dans ce travaille les propriétés asymptotiques et l'estimation des paramètres pour les processus $ARCH$ périodiquement corrélés (PC). Nous donnons la définition d'un processus $PCARCH\,(q)$ et quelques propriétés probabilistes en basant sur les représentations $BEKK$ et VEC. Nous établissons les propriétés asymptotiques de la moyenne empirique et la covariance.

Abstract

This paper studies the asymptotic properties of first and second empirical moments for squared causal periodically correlated (PC) $ARCH$ processes. The definition of $PCARCH\,(q)$ processes is first given and some probabilistic properties based on representations $BEKK$ and VEC are discussed.
Mots clés. Processus $ARCH$ périodiques; Processus périodiquement corrélés; Normalité asymptotique; Consistance.

1 Introduction

Les processus $(G)ARCH$ périodiques ont eu un grand succès en économétrie et en statistique. Cette classe de processus a été initiée par Bollerslev et Ghysels [4] et continue à gagner en popularité, particulièrement pour analyser les séries financières. Ces processus sont semblables aux modèles de Markov-switching $GARCH$ ($MS-GARCH$) (voir par exemple [6]) mais la différence majeure est que la suite qui gouverne les paramètres du modèle n'est pas la réalisation d'une chaîne mais est une fonction périodique du temps. Cette spécification rend le processus non stationnaire globalement, mais il est stationnaire dans chaque période. Plusieurs articles sont consacrés aux propriétés asymptotiques des modèles $P-GARCH$. Aknouche et Bibi [1] et Bibi et Lescheb [3] sont des références traitant de l'estimation des coefficients périodiques dans les modèles $P-(G)ARCH$ et son comportement asymptotique. Dans ce papier, nous considérons le processus $ARCH$ périodique (Autorégressifs Conditionnellement Hétéroscédastiques) noté $PARCH$ introduit premièrement par Bollerslev et Ghysels [4] comme un meilleur moyen caractérisant les modèles périodiques ou saisonniers dans la volatilité du marché financier. Un processus

de second ordre $(X_t)_{t\in\mathbb{Z}}$ est dit $ARCH(q)$ périodique ($PARCH(q)$) avec période $s \geq 1$, s'il satisfait l'équation

$$X_t(v) = e_t(v)\, h_t(v) \;\; et \;\; h_t^2(v) = w(v) + \sum_{k=1}^{q} \alpha_k(v)\, X_t^2(v-k). \tag{1}$$

Dans l'équation (1) $X_t(v)$ (respectivement $h_t(v), e_t(v)$) se reporte à X_t (respectivement h_t, e_t) pendant la v-ième "saison" ou "régime", $1 \leq v \leq s$ de la période t, c'est-à-dire, $X_t(v) := X_{st+v}$ (respectivement $h_t(v) := h_{st+v}, e_t(v) := e_{st+v}$) et la commodité $X_t(v) := X_{t-1}(s-v)$ (respectivement $h_t(v) := h_{t-1}(s-v)$ et $e_t(v) := e_{t-1}(s-v)$) si $v \leq 0$. Ce type de modèle est (en général) non stationnaire (voir [2]). Sous des conditions appropriées qui seront considérées ultérieurement, ce type de processus appartient à la classe des processus périodiquement corrélés (PC) caractérisés par $E\{X_t\} = E\{X_{t+s}\}$ et $Cov(X_t, X_r) = Cov(X_{t+s}, X_{r+s})$ pour tous t et r. Notons que si $s = 1$, un processus PC est réduit à un processus stationnaire au second ordre. Si $s > 1$, le processus est globalement non stationnaire, mais il est stationnaire dans chaque période. Lorsque en estimant ce modèle, nous supposons que (i) $w(v) > 0$ pour tout $v \in \{1, ..., s\}$. (ii) $\alpha_k(v) \geq 0$ pour tous $v \in \{1, ..., s\}$ et $k = 1, ..., q$. (iii) e_t est une suite de variables aléatoires $i.i.d$ avec $E\{e_t\} = 0, \sigma_e^2 = E\{e_t^2\} = 1$ et le moment d'ordre quatre κ_4 est fini où $\kappa_n = E\{e_t^{2n}\}, n \geq 2$.

Les suppositions (i) et (ii) assurent la positivité forte de $h_t^2(v)$ ces conditions peuvent être formulées d'après Nelson et Cao [7] qui suggèrent la condition de restreindre les deux coefficients saisonniers pour être non négative avec le coefficient de l'interception saisonnier strictement positif.

2 Représentations vectorielles et ses propriétés

C'est bien clair qu'il y a une équivalence alternative entre la stationnarité multicritère et le processus PC (voir [8]). Plus précisément $\underline{e}_t = (e_t(1), ..., e_t(s))', \underline{w} = (w(1), ..., w(s))'$, $\underline{h}_t = (h_t(1), ..., h_t(s))'$ et $\underline{X}_t = (X_t(1), ..., X_t(s))'$ est un processus $ARCH$ multivariate sur la forme $BEKK$, c'est-à-dire, $\underline{X}_t = H_t \underline{e}_t \; et \; H_t^2 = w + C_0 + \sum_{k=1}^{q} C_k' \underline{X}_{t-k} \underline{X}_{t-k}' C_k$, ou sur la forme VEC, c'est-à-dire, $\underline{X}_t = diag(\underline{h}_t)\, \underline{e}_t \; et \; \underline{h}_t^2 = \underline{w} + C_0 \underline{\mathbf{1}} + \sum_{k=1}^{q} A_k \underline{X}_{t-k}^2$ où $w = diag(\underline{w}), H_t = diag(\underline{h}_t), A_k' = C_k^{\odot 2}$ avec $(C_0)_{k,j} = \left(\sum_{k=1}^{q} \alpha_k(1)\, X_t^2(s-k+1) \right) \mathbb{I}_{\{i=1,j=1\}}$, $\underline{X}_{t-k} = (X_t(2-k), X_t(3-k), ..., X_t(s-k+1))', (C_k)_{i,j} = \sqrt{\alpha_k(j)}\mathbb{I}_{\{i=j-1,j=2,...,s\}}, i = 1, ..., s, k = 1, ..., q$ et $X_t(v) X_t(v') = 0$ pour $v \neq v', v, v' \in \{1, ..., s\}$. Puisque $\underline{X}_t^2 =$

$diag\left(\underline{e}_t^2\right)\underline{h}_t^2$, il est facile de vérifier que $\underline{X}_t^2$ admet la représentation suivante

$$\underline{X}_t^2 = \sum_{k=1}^{q} \Phi_k\left(\underline{e}_t\right)\underline{X}_{t-k}^2 + \underline{\eta}\left(\underline{e}_t\right) \tag{2}$$

où $\left(\Phi_k\left(\underline{e}_t\right)\right)_{i,j} = e_t^2\left(i\right)\alpha_k\left(i\right)\mathbb{I}_{\{i=j+1,j=1,...,i\}}, i = 1,...,s, k = 1,...,q$, $\Phi_k\left(\underline{e}_t\right) = O_{(s)}$ pour $k > q$ et

$$\underline{\eta}\left(\underline{e}_t\right) = \left(e_t^2\left(1\right)\left(w\left(1\right) + \sum_{k=1}^{q}\alpha_k\left(1\right)X_t^2\left(1-k\right)\right), e_t^2\left(2\right)w\left(2\right), ..., e_t^2\left(s\right)w\left(s\right)\right)'$$

avec matrice de variance covariance Σ_η. La représentation (2) est équivalente à

$$\underline{x}_t^2 = M_t\underline{x}_{t-1}^2 + \underline{\eta}_t \tag{3}$$

où $\underline{x}_t^2 = \left(\underline{X}_t^{2\prime}, \underline{X}_{t-1}^{2\prime}, \underline{X}_{t-2}^{2\prime}, ..., \underline{X}_{t-q+1}^{2\prime}\right)'$, $\underline{\eta}_t = \left(\underline{\eta}'\left(\underline{e}_t\right), \underline{O}_{(s)}', ..., \underline{O}_{(s)}'\right)'$ et la matrice M_t est donnée par $\left(M_t\right)_{i,j} = \left(\Phi_j\left(\underline{e}_t\right)\mathbb{I}_{\{i=1\}} + I_{(s)}\mathbb{I}_{\{i=j+1,j\}}\right)$.

Dans ce qui suit, si le processes $\left(X_t\right)_{t\in\mathbb{Z}}$ est un processus PC strictement stationnaire causal et ergodique alors le vecteur associé $\left(\underline{X}_t\right)_{t\in\mathbb{Z}}$ est strictement stationnaire, stationnaire au second ordre, causal et ergodique. Puisque le processus $\left(M_t, \underline{\eta}_t\right)_{t\in\mathbb{Z}}$ est un couple des matrices et des vecteurs aléatoires indépendents et conjointement identiquement distribués, indépendant de $\underline{X}_s$ pour $s < t$ et si $E\left\{\log^+\left\|M_0\right\|\right\}$ est finie, où $\log^+\left|x\right| = \max\left\{\log\left|x\right|, 0\right\}$, alors, suivant Bougerol et Picard [5], l'unique solution causale, strictement stationnaire et ergodique de l'équation (3) est donnée par

$$\underline{x}_t^2 = \sum_{k=0}^{\infty}\left\{\prod_{j=0}^{k-1}M_{t-j}\right\}\underline{\eta}_{t-k} + \underline{\eta}_t,\ t\in\mathbb{Z} \tag{4}$$

si l'exposant de Lyapunov définit par $\delta_L = \inf_{t>0}\left\{\frac{1}{t}E\log\left\|\prod_{j=0}^{t-1}M_{t-j}\right\|\right\}$ est strictement négatif, cependant la série (4) converge $p.s.$ Cette condition n'a pas beaucoup d'intérêt pratique car elle dépend de calcul de produit infini des matrices aléatoires, ce qui rend délicat d'effectuer les calculs, d'après la définition de δ_L, on a

C.1 $\delta_L \leq E\left\{\log\left\|M_0M_1\right\|\right\} < 0$.

Donc cette condition pourrait être employée comme étant une condition suffisante pour l'existence d'une solution strictement stationnaire

Pour rendre une théorie de l'estimation possible, la solution du processus doit avoir quelques moments, cependant le critère de Lyapunov ne garantit pas l'existence de tels moments. Nous devons par conséquent, chercher des conditions assurent l'existence des moments pour une solution strictement stationnaire lesquelles, l'exposant de Lyapunov sera automatiquement négatif.

Théorème 1 *Soient $(X_t)_{t\in\mathbb{Z}}$ un processus définit par (1) et $(\underline{x}_t^2)_{t\in\mathbb{Z}}$ sa représentation vectorielle, alors l'équation (3) a une solution strictement stationnaire dans $\mathbb{L}_1$ si et seulement si la condition **C. 1** est vérifiée. De plus, la solution est unique, causale, ergodique et donnée par (4) laquelle converge p.s et dans $\mathbb{L}_1$.*

Corollaire 2 *Une solution PC stationnaire au sens strict et au second ordre, causale, ergodique avec période s donnée par $X_t(v) = e_t(v) h_t(v)$ et $h_t^2(v) = \sum_{j\geq 1}\left\{\prod_{i=1}^{j-1}\alpha_i(v-i)\, e_t^2(v-i-1)\right\} w(v-j)$ existe à (1) quand la condition **C. 1** est vérifiée, où $\max_{1\leq v\leq s}\sum_{j\geq 1}\left\{\prod_{i=1}^{j-1}\alpha_i(v-i)\right\} w(v-j) < \infty$.*

Théorème 3 *Soient le processus $(X_t)_{t\in\mathbb{Z}}$ généré par (1) et $(\underline{x}_t^2)_{t\in\mathbb{Z}}$ sa représentation vectorielle laquelle satisfait là 'équation (3). Alors pour tout $m \geq 1$, les deux relations suivantes sont équivalentes*

1. $E\{\underline{x}_t^{2m}\} < \infty$
2. $\rho\left(M^{(m)}\right) < 1$, *où* $M^{(m)} = E\left\{M_0^{\odot m}\right\}$ *si* $m > 1$ *et* $M = E\{M_0\}$.

3 Structure de la covariance

Par conséquence, la solution PC $(X_t)_{t\in\mathbb{Z}}$ dans $\mathbb{L}_2$ est un bruit blanc faible. Pour distinguer entre ceux-ci et les modèles $ARCH$, nous avons besoin d'enquêter le comportement de quelques moments d'ordre supérieurs à 2, dans ce sens, nous considérons dans la suite l'analyse des fonctions de covariances saisonnières à h, c'est-à-dire, $\gamma_v(h) = Cov\left(X_t^2(v), X_t^2(v-h)\right)$ qui prend en considération la stationnarité au second ordre du processus $\left(\underline{X}_t^2\right)_{t\in\mathbb{Z}}$. Nous dérivons en premier pas la moyenne saisonnière du modèle $PARCH(q)$. Posons $\mu_v = E\{X_t^2(v)\}$ on a $\mu_v = w(v) + \sum_{k=1}^{q}\alpha_k(v)\mu_{v-k}$. On peut réécrire cette équation sous la représentation vectorielle suivante $\underline{\mu}_v = A(v)\underline{\mu}_{v-1} + \underline{\epsilon}(v)$ où $\underline{\mu}_v = \left(\mu_v, ..., \mu_{v-q+1}\right)'$, $\underline{\epsilon} = (w(v), 0, ..., 0)'$ et la matrice $A(v)$ est définie par $(A(v))_{i,j} = \alpha_j(v)\mathbb{I}_{\{i=1\}} + 1\mathbb{I}_{\{i=j+1,j\}}$, $j = 1, ..., q$. Par récurrence s fois l'équation précédente et en nécessitant que $\underline{\mu}_0 = \underline{\mu}_s$ on trouve $\underline{\mu}_s = \left(I_{(q)} - \prod_{v=1}^{s}A(v)\right)^{-1}\sum_{j=1}^{s}\left(\prod_{v=j+1}^{s}A(v)\right)\underline{\epsilon}(j)$. De plus $\underline{\mu}_v = \left\{\prod_{j=1}^{v}A(j)\right\}\underline{\mu}_s + \sum_{j=1}^{v}\left(\prod_{l=j+1}^{v}A(l)\right)\underline{\epsilon}(j)$, $v = 1, ..., s-1$. Maintenant, posons $\eta_t(v) = X_t^2(v) - h_t^2(v)$ alors d'après (1) nous avons $X_t^2(v) = w(v) + \sum_{k=1}^{q}\alpha_k(v)X_t^2(v-k) + \eta_t(v)$, en considérant le processus centré $X_t^2(v) - \mu_v$ dans la représentation PAR nous obtenons

$X_t^2(v) - \mu_v = \sum_{k=1}^{q} \alpha_k(v)\left(X_t^2(v-k) - \mu_{v-k}\right) + \eta_t(v)$. Alors, en multipliant les deux membres de cette équation par $X_t^2(v-h), h \geq 0$ et en appliquant l'espérance nous obtenons $\gamma_v(h) = \sigma^2(v)\mathbb{I}_{\{h=0\}} + \sum_{k=1}^{q} \alpha_k(v)\gamma_{v-k}(h-k)\mathbb{I}_{\{h\geq 1\}}$, cette équation est équivalente à

$$\underline{\gamma}_v(h) = A(v)\underline{\gamma}_{v-1}(h) + \underline{\sigma}$$

où $\underline{\gamma}_v(h) = \left(\gamma_v(h), ..., \gamma_{v-q+1}(h)\right)'$,
$\underline{\gamma}_{v-1}(h) = \left(\gamma_{v-1}(h)\mathbb{I}_{\{h\geq 1\}}, ..., \gamma_{v-q}(h)\mathbb{I}_{\{h\geq 1\}}\right)', \underline{\sigma} = \left(\sigma^2(v)\mathbb{I}_{\{h=0\}}, 0, ..., 0\right)'$. Rappelons que l'équation (2) peut être s'écrit sous la forme

$$\underline{X}_t^2(v) = A(v)\underline{X}_t^2(v-1) + \underline{\delta}(v), v = 1, ..., s$$

où $\underline{X}_t^2(v) = \left(e_t^2(v)X_t^2(v), ..., e_t^2(v-q+1)X_t^2(v-q+1)\right)', \underline{\delta}(v) = \left(e_t^2(v)w(v), 0, ..., 0\right)'$,
$\underline{X}_t^2(v-1) = \left(e_t^2(v-1)X_t^2(v-1), ..., e_t^2(v-q)X_t^2(v-q)\right)'$, ce qui est équivalent à

$$\underline{y}_t^2 = A\underline{y}_{t-1}^2 + \underline{\delta}$$

où $\underline{y}_t^2 = \left(\underline{X}_t^{2\prime}(1), \underline{X}_t^{2\prime}(2), ..., \underline{X}_t^{2\prime}(s)\right)', \underline{y}_{t-1}^2 = \left(\underline{X}_t^{2\prime}(0), \underline{X}_t^{2\prime}(1), ..., \underline{X}_t^{2\prime}(s-1)\right)'$,
$\underline{\delta} = \left(\underline{\delta}'(1), ..., \underline{\delta}'(s)\right)'$ et $A = diag\{A(1), ..., A(s)\}$.

3.1 Propriétés asymptotiques de la moyenne empirique et la covariance

Soit $\{X_1, ..., X_n\}$ une réalisation de longueur $n = sN$ de l'unique solution PC pour le modèle (1). Alors le problème est équivalent d'avoir une réalisation $\{\underline{X}_1, ..., \underline{X}_N\}$ d'un processus $(\underline{X}_t)_{t\in\mathbb{Z}}$ stationnaire au second ordre. Premièrement, définissons des estimateurs pour μ_v et pour $\gamma_v(h)$ comme suit $\widehat{\mu}_v := \frac{1}{N}\sum_{t=0}^{N-1} X_t^2(v), \widehat{\gamma}_v(h) := \frac{1}{N}\sum_{t=0}^{N-1} X_t^2(v)X_t^2(v-h) - \widehat{\mu}_v\widehat{\mu}_{v-h}$ pour tout $v \in \{1, ..., s\}$ et pour tout $h \geq 0$. Deuxièment, définissons les vecteurs $\underline{\widehat{\mu}} := \left(\underline{\widehat{\mu}}_1, ..., \underline{\widehat{\mu}}_s\right), \underline{\mu} := \left(\underline{\mu}_1, ..., \underline{\mu}_s\right), \underline{\widehat{\gamma}}(h) := \left(\underline{\widehat{\gamma}}_1(h), ..., \underline{\widehat{\gamma}}_s(h)\right)$ et $\underline{\gamma}(h) := \left(\underline{\gamma}_1(h), ..., \underline{\gamma}_s(h)\right)$. Le résultat suivant caractérise le comportement asymptotique de la moyenne empirique.

Proposition 4 *Considérons le processus PARCH et soit* $\left(\underline{y}_t^2\right)_{t\in\mathbb{Z}}$ *sa représentation vectorielle associée. Si* $\left(\underline{y}_t^2\right)_{t\in\mathbb{Z}}$ *admet des moments d'ordre supérieurs à 2, alors pour tous* $v, v' \in \{1, ..., s\}$

1. $\underline{\widehat{\mu}}_v$ *converge vers* $\underline{\widehat{\mu}}$ *p.s.*
2. $\lim_{N\to\infty} NCov\left(\underline{\widehat{\mu}}_v, \underline{\widehat{\mu}}_{v'}\right) = (V_{as})_{v,v'} := \sum_{k\in\mathbb{Z}} \underline{\gamma}_{v'}(v'-v+sk)$ *où* $V_{as} := \sum_{h\in\mathbb{Z}} Cov\left(\underline{y}_t^2, \underline{y}_{t-h}^2\right)$

3. $\lim_{N\to\infty} E\left\{\underline{\widehat{\mu}}_v - \underline{\mu}_v\right\}^2 = \underline{O}_{(q)}$

4. *Le vecteur* $\sqrt{N}\left(\underline{\widehat{\mu}} - \underline{\mu}\right)$ *converge en loi vers* $\mathcal{N}(0, V_{as})$.

Le résultat suivant caractérise le comportement asymptotique de la matrice de covariance $\underline{\widehat{\gamma}}_v(h)$.

Proposition 5 *Considérons le carré du processus PARCH et soit* $\left(\underline{y}_t^2\right)_{t\in\mathbb{Z}}$ *sa représentation vectorielle associée. Si* $\left(\underline{y}_t^2\right)_{t\in\mathbb{Z}}$ *admet des moments d'ordre supérieurs à 4, alors pour tous* $h \geq 0$ *et* $v, v' \in \{1, ..., s\}$

1. $\underline{\widehat{\gamma}}_v(h)$ *converge vers* $\underline{\gamma}_v(h)$ *p.s*

2. $\lim_{N\to\infty} NCov\left(\underline{\widehat{\gamma}}_v(h), \underline{\widehat{\gamma}}_{v'}(k)\right) = \left(W_{as}(h,k)\right)_{v,v'}$
   où $W_{as}(h,k) := \sum_{l\in\mathbb{Z}} Cov\left(\underline{y}_t^2 \odot \underline{y}_t^2(h), \underline{y}_{t-l}^2 \odot \underline{y}_{t-l}^2(k)\right)$

3. $\lim_{N\to\infty} E\left\{\underline{\widehat{\gamma}}_v(h) - \underline{\gamma}_v(h)\right\}^2 = \underline{O}_{(q)}$

4. *Le vecteur* $\sqrt{N}\left(\underline{\widehat{\gamma}}(h) - \underline{\gamma}(h)\right)$ *converge en loi vers* $\mathcal{N}(0, W_{as}(h,h))$.

Bibliographie

[1] Aknouche, A., and A., Bibi (2009) Quasi-maximum likelihood estimation of periodic $GARCH$ and periodic $ARMA - GARCH$ processes. *J. Time Ser. Anal.* 30 (1) pp 19–46.

[2] Bibi, A. and C. Francq (2003). Consistent and asymptotically normal estimators for cyclically time-dependent linear models. *Ann. Inst. Statist. Math.* 55,1-13.

[3] Bibi, A., and I., Lescheb (2011) Estimation and asymptotic inference in first order periodic $(I)GARCH$ models. *Communication in Statistics: Theory and Methods.* To appear.

[4] Bollerslev, T., and E., Ghysels (1996). Periodic autoregressive conditional heteroskedasticity. *J. of Business & Economic Statistics*, 14, pp. 139-151.

[5] Bougerol, P. & N., Picard (1992). Stationarity of GARCH processes and some nonnegative time series. *Journal of Econometrics*, 52, pp. 115-127.

[6] Francq, C., and J-M., Zakoîan (2005) The L^2-Structures of standard and switching-regime $GARCH$ models. *Stoch. Processes and their App.* Vol. 115, 1557-1582.

[7] Nelson, D. R., and C. Q., Cao (1992). Inequality constraints in the univariate GARCH model. *J.B.E.S.* Vol. 10, pp. 229-235.

[8] Tiao,G. C., & M. R., Grupe (1980). Hidden periodic autoregressive-moving average models in time series data. *Biometrika*, 67, pp. 365-373.

An optimal confidence interval for an adjusted premium estimator in simulated insurance data

Kmar Fersi*, Kamel Boukhetala**
Samir Ben Ammou***

*Institut Supérieur de Gestion Sousse,
Computational Mathematics Laboratory, Route de Kairouan, 5019 Monastir, Tunisia
fersi.gmar@yahoo.fr
** Faculté de Mathématiques Bp. 32, El-Alia, Bab-Ezzouar, USTHB Alger, Algeria
*** Computational Mathematics Laboratory, Route de Kairouan, 5019 Monastir, Tunisia

Résumé. L'objectif de ce travail est de déterminer un intervalle de confiance optimal pour l'estimateur de la prime ajustée, développé par [7] en assurance non vie. Pour cela, on est amené à améliorer la stratégie de minimisation de la variance de cet estimateur sous contraintes stochastiques liées aux excès des sinistres au-delà d'un seuil u ([4]). Cependant, la qualité des résultats de minimisation de ce problème stochastique, obtenue par l'Algorithme Génétique recuit Simulé (AGS) ([6] et [10]), est très sensible au nombre des risques extrêmes. Lorsque la proportion des excédants est faible, une technique de simulation, basée sur des observations réelles est proposée, afin de déterminer une taille d'échantillon simulé adéquat au calcul d'un intervalle de confiance optimal pour cet estimateur.

Mots clés: Intervalle de confiance, AGS, prime ajustée, simulation Monte Carlo, risques extrêmes, assurance non-vie.

1 Introduction

Les limites de la "solvabilité I" ont accéléré l'adoption de la directive "solvabilité II" par la commission Européenne. En effet, cette directive encourage les compagnies à opter pour un modèle interne afin qu'elles soient en mesure par elles-mêmes d'apprécier et de mesurer leurs risques. L'approche par le modèle interne est la seule à apporter des éléments permettant une meilleure maîtrise des risques et de dégager une alternance de stratégie de tarification spécifique à la réassurance qui permet de mutualiser à l'échelle mondiale les risques des assureurs et réduire la sévérité des sinistres extrêmes [2].
D'un point de vue économique, un risque est considéré comme assurable si, selon les critères de [1], il existe un transfert de risque entre l'assuré et l'assureur qui soit mutuellement avantageux pour les deux parties. Compte tenu de la sinistralité historique en assurance non-vie et les actions de prévention qui y sont réalisées, nous avons retenu l'hypothèse, contestable, que les ménages surestiment les événements rares et sous-estiment les événements fréquents [4].

En effet, les assurances dommages sont fondées sur une mutualisation des risques [11], segmentés et sélectionnés à l'aide d'outils statistiques contribuant à définir les bons et les mauvais risques et à établir une tarification adéquate aux clients. Mais, lorsque la proportion de risques extrêmes est faible, la diversité de produits d'assurance devient susceptible de présenter des difficultés à la stratégie de réduction de la variance de l'estimateur de la prime ajustée ([4]). Dans ce travail, On propose une amélioration de cette stratégie, à l'aide des techniques de simulation basées sur des observations réelles, afin de déterminer un intervalle de confiance optimal pour cet estimateur.

2 Simulation de Monte- Carlo et réduction de la variance de la prime ajustée

Les méthodes de simulation Monte- Carlo sont certainement les techniques de simulation les plus répandues. Ce sont celles que nous utiliserons pour étudier la performance de la stratégie de réduction de la variance pour différentes tailles des échantillons des excédants simulés selon la loi de Pareto généralisée (Generalized Pareto Distribution ou GPD). Les paramètres de cette loi sont estimés à partir d'un échantillon de données réelles de 50000 observations pour des véhicules 4 roues de tourisme durant l'année 2004, issus du portefeuille d'une mutuelle d'assurance Française.
L'intérêt de cette approche de simulation est de reproduire ce qui se passe dans le monde réel de manière fidèle et avec un niveau de significativité statistiquement appréciable. On utilise un mécanisme qui génère les données de façon approximativement similaire que celui qui opère dans le monde réel des sinistres extrêmes.

Ajustement statistique des sinistres extrêmes en assurance automobile

Considérons l'exemple des coûts de sinistres d'un portefeuille d'assurance française. La détermination du seuil par la méthode Peak Over Threshold (P.O.T) ([5]) permet de calculer une prévision du coût d'un sinistre extrême pour une probabilité d'occurrence de 99,9% d'être une valeur extrême et avec un niveau de confiance de 95%.
La figure (1) présente une estimation ponctuelle du quantile extrême ainsi que la limite inférieure et supérieure de la perte maximale pour la compagnie d'assurance en cas de survenance d'un sinistre grave avec une probabilité de 0,1%.
En effet, la méthode P.O.T détecte toujours moins de valeurs extrêmes que des autres méthodes classiques. Plus le seuil est grand, plus on s'intéresse aux événements extrêmes mais comme la prime ajustée sera estimée avec peu de données, sa variance sera grand. Le système de tarification des sinistres très élevés semble toutefois lacunaire lorsque la proportion des excédants est faible. Cependant, on utilise des techniques de simulation de Monte-Carlo basées sur des données réelles pour réaliser un échantillon des excédants simulés de taille adéquate à la détermination d'un intervalle de confiance optimal pour l'estimateur de la prime ajustée.

Amélioration de l'efficacité des techniques de réduction de la variance

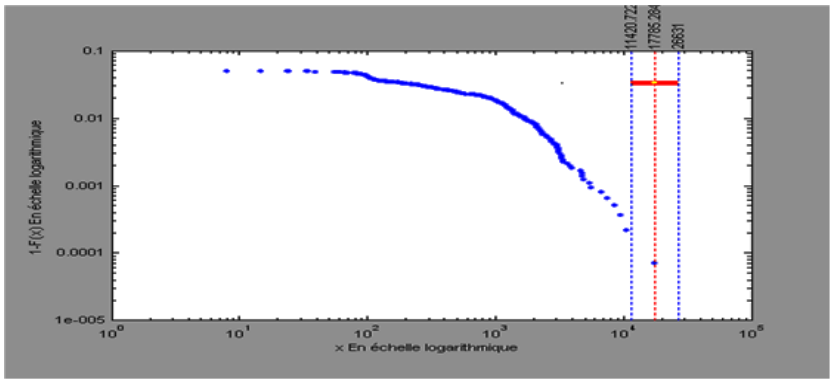


FIG. 1 – *La prévision d'un coût extrême*

L'objectif de notre travail est de montrer comment la méthode de simulation de Monte Carlo permet de repousser les limites de l'insuffisance du nombre de risques extrêmes et d'améliorer la stratégie développée par [4]. Une telle stratégie permet de réduire la variance $\sigma^2(p,\xi,\beta)$ de l'estimateur de la prime ajustée sous contraintes stochastiques ; Ce type de problème est généralement difficile à résoudre. On propose d'utiliser l'algorithme génétique recuit simulé (AGS) ([6] et [10]) pour chercher un optimum global à ce problème qui est présenté comme suit :

$$(Q) = \begin{cases} \min_{\xi,\beta,p} \sigma^2(\xi,\beta,p) = \frac{(1+\xi)\beta^2}{(1/p-\xi)^2}\left[\frac{(1/p-1)^2}{1+\xi} + \frac{1+\xi}{(1/p-\xi)^2} - \frac{2}{1/p-\xi} + 2\right] \\ \frac{1}{k}\sum_{i=1}^{k} \log(1+\xi\frac{y_i}{\beta}) = \xi \qquad (1) \\ \frac{1}{k}\sum_{i=1}^{k} \log\frac{y_i/\beta}{1+y_i/\beta} = \frac{1}{1+\xi} \qquad (2) \\ p = \psi_\alpha(x) \\ \frac{1}{2} \leq \xi < \frac{1}{p},\ 2 > p \geq 1,\ \beta > 0 \end{cases}$$

Avec $y_1, ..., y_k$: sont les x_i observations réelles dépassant le seuil u,
(1) et (2) : issuent de l'estimation par la méthode du maximum de vraisemblance des paramètres de GPD suggérés par [8]
p : est un paramètre de distortion selon le principe de [9]
$\psi_\alpha(z) = 1 + z^\alpha$, où z est une variable aléatoire uniformément distribuée sur l'intervalle $[0,1]$ et $\alpha > 0$
La minimisation de ce problème par l'AGS est très sensible au choix de k_i. On étudie ensuite le comportement asymptotique des résultats de chaque scénario des excédants simulés.

Simulation et interprétation des résultats de l'AGS

Dans ce cas, le nombre des excès est déterministe. Trois valeurs différentes de $k_i; i = \{40, 80, 120\}$ ont été prises dans le but d'étudier l'impact de ce facteur.
Cependant, pour permettre à l'actuaire d'effectuer un choix optimal de la taille de l'échantillon des excès simulés, il est important de vérifier l'évolution de la convergence vers l'optimum glo-

bal pour une vaste bibliothèque de scénarios de différentes tailles.
Il semble raisonnable de justifier, par la figure (2), que l'amélioration de la convergence

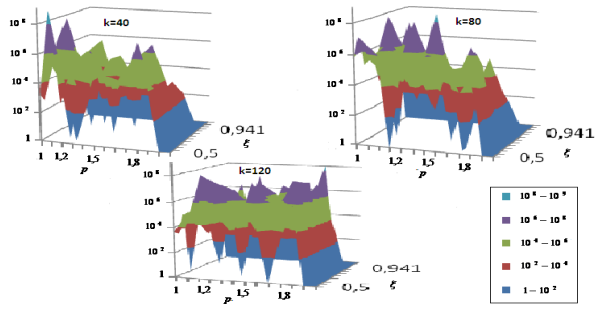


FIG. 2 – *Courbe des configurations optimales de la variance de l'estimateur de la prime ajustée pour des scénarios d'échantillons simulés de différentes tailles* k_i

n'est pas grossièrement fausse, lorsque le nombre des excédants simulés devient suffisamment grand.
Un tel choix peut se limiter à un examen visuel de la suite de plusieurs exécutions par l'AGS ou faire en sus appel à des tests d'ajustements graphiques.
La figure (3) permet de souligner que la normalité de l'ajustement des résultats de la variance,

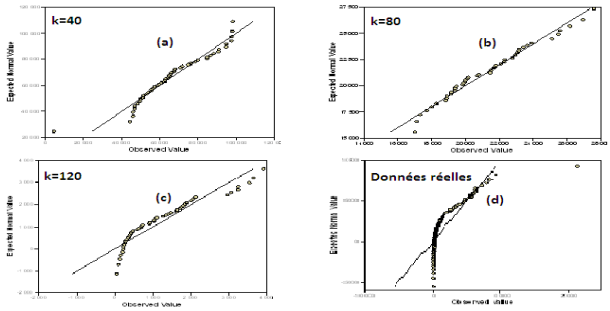


FIG. 3 – *Distribution Box Plot de la variance de l'estimateur de la prime ajustée pour différentes tailles* k_i *de sinistres extrêmes simulés et des excédants réels d'un portefeuille d'assurance automobile*

obtenus par l'AGS, pour les données réelles (figure (d)), s'améliore par l'utilisation de cette technique de simulation. En conséquence, on peut noter que les représentations graphiques pour les scénarios simulés (figure (a),(b) et (c)) justifient clairement les résultats théoriques obtenus par [7]. Il n'est pas donc étonnant de constater une minimisation de l'amplitude de l'intervalle de confiance de la prime ajustée. Ainsi pour l'assurance, plus le nombre des contrats agrégés est important, plus il autorise, toutes choses égales par ailleurs, une perspective d'établir des niveaux de primes de risques raisonnables.

Conclusion

L'augmentation du nombre des valeurs extrêmes par les techniques de simulations de Monte Carlo semble être une stratégie avantageuse pour la problématique de la réduction de la variance de l'estimateur de la prime ajustée. cette approche permet également d'améliorer la normalité de l'ajustement des résultats optimaux obtenus par l'algorithme génétique recuit simulé. Cependant, l'étude du comportement asymptotique de ces résultats justifie les résultats théoriques obtenus par [7].

Nous avons également relevé que les règles de tarification des risques extrêmes peuvent passer par une modernisation des outils de calcul de la prime pure. De plus, ces techniques stratégiques provoquent une diversification des moyens d'indemnisation de manière à ajuster au mieux la tarification et la prise en compte de l'efficacité de la mutualisation pour des portefeuilles homogènes.

Références

[1] Berliner, B. (1982). Limits of insurability of risks. Prentice Hall, Swiss Re Economic Research and Consulting.

[2] Centeno, M.L.& Guerra, M. (2010). The optimal reinsurance strategy : the individual claim case. Insurance : Mathematics and Economics, Vol. 46, p.450-460

[3] Denuit M. & Charpentier A. (2005). Mathématiques de l'assurance non-vie : Tarication et provisionnement. Economica, Paris

[4] Fersi K., Boukhetala K. & Ben Ammou S. (submitted (2010)). Optimal strategy for reduction variance of estimator of the extreme risks adjusted premium. Application in car insurance data. European Actuarial Journal

[5] Fersi K., Benlagha N. & Ben Ammou S. (submitted (2010)). The insurability of risks : A quantitative approach applied to motor insurance. Scandinavian Actuarial Journal

[6] Kirkpatrick S. & al. (1983). Optimization by simulated annealing. Science, Vol. 220, p.671-680.

[7] Necir A. & Boukhetala K. (2004). Estimating the risk-adjusted premium for the largest claims reinsurance covers. COMPSTAT'2004. Proceeding in Computational Statistics. ISBN 3-7908-1554-3, Physica-Verlag, Heidelberg, New York. Springer

[8] Pickands, J. I. (1975). Statistical inference using extreme value order statistics. Annals of Statististics, Vol. 3, p. 119-131.

[9] Wang S. (1996). Premium calculation by transforming the layer premium density. ASTIN Bulletin, Vol. 26, p.71-92

[10] Youssef H., Sadiq M.S. & Adiche H. (2001). Evolutionary algorithms, simulated annealing and tabu search : a comparative study, Engng. Appl. Artificial Intellingence, Vol. 14, p.167-181

[11] Zajdenweber D. (2006). Economie et gestion de l'assurance, Economica

Summary

The aim of this work is to determine an optimal confidence interval for the adjusted premium estimator, developed by [7] in non-life insurance. For this, we are led to improve the minimization variance strategy of this estimator under stochastic constraints related to excesses over thresholds u ([4]). The quality of the results of this stochastic minimization problem, obtained by Genetic Simulated annealing Algorithm (GSA) ([6] and [10]), is very sensitive to the number of extreme risks. When the proportion of the excess is low, a simulation technique based on actual observations is proposed to determine an adequate size of simulated sample likely to calculate an optimal confidence interval for the estimator.

Keywords: Confidence interval, GSA, adjusted premium, Monte Carlo simulation, extreme risks, non-life insurance.

Testing Scale Efficiency: A Smooth Bootstrap Approach

Hédi Essid,[(i)] Pierre Ouellette[(ii)] & Stéphane Vigeant[(iii)]

[(i)] Institut Supérieur de Gestion, Université de Tunis (Corresponding author)
41, Rue de la liberté, Cité Bouchoucha 2000, Le Bardo, Tunis-Tunisie.
e-mail: hedi.essid@ihec.rnu.tn

[(ii)] Dept. of Economics, Université du Québec à Montréal, Canada.
e-mail: ouellette.pierre@uqam.ca

[(iii)] IESEG School of management and EQUIPPE (Universités de Lille), UST de Lille, France.
e-mail: Stephane.Vigeant@univ-lille1.fr

Résumé : Ce papier présente une procédure de test statistique non paramétrique pour mesurer l'efficacité à l'échelle d'une organisation. Cette procédure permet au praticien de vérifier si le score d'efficacité observé est réel ou bien due à la variation échantillonnale. Nous utilisons la méthode d'enveloppement des données (DEA) pour estimer les scores d'efficacité et nous adoptons une méthodologie bootstrap lisse homogène pour approximer la distribution d'échantillonnage de la statistique du test. Les résultats de simulation Monte Carlo témoigne de la performance de notre procédure de test.

Summary: This paper presents a nonparametric statistical test procedure for organization scale efficiency. This procedure allows the practitioner to test whether the observed scale efficiency score is real or due to sampling variation. We use the Data Envelopment Analysis method (DEA) to estimate efficiency scores and we adopt a smooth homogeneous bootstrap methodology to approximate the sampling distribution of the test statistic. The results of Monte Carlo simulation show the performance of our testing procedure.
Keywords: Test of Hypothesis, Data envelopment analysis, Kernel, Bootstrap, Monte Carlo.

1. Introduction

In this paper, we address the question of the qualitative measurement of returns to scale in organizations. This measure allows us to determine if organizations' activities are characterized by either increasing, decreasing or constant returns to scale. We develop a procedure for testing statistical hypotheses based on efficiency scores of Färe and Grosskopf (1988). We estimate these efficiency scores by the DEA method. Then we use a smooth homogeneous bootstrap method for approximating the sampling distribution of the estimators of these scores. To investigate the performance of our testing procedure, we conduct a study of Monte Carlo simulation.

2. Production technology and returns to scale

Consider a production activity that uses the sets of inputs, $x=\{x_i, i=1,\ldots,m\}$ to produce the output vector, $y=\{y_r, r=1,\ldots,s\}$. The production possibility set of this activity is defined as:[1]

$$\Psi=\left\{(x,y)\in\mathbb{R}_+^{m+s}\,\middle|\,(x,y)\text{ is feasible}\right\}. \quad (1)$$

[1] We suppose that the production set Ψ is closed, satisfies free disposal of inputs, is bounded for finite inputs and positive outputs require positive inputs. These are standard assumptions on the technology and are discussed in Färe (1988), among others.

The returns to scale of the firms are characterized by the way one can either expand the scale in the production set or shrink it or both. That is, the technology exhibits non increasing (*nirs*), non decreasing (*ndrs*) or constant (*crs*) returns to scale depending on the value assumed by the positive scalar α in the following set:

$$\Psi^k = \left\{(x,y) \in \Psi \middle| (\alpha x, \alpha y) \in \Psi \text{ for all } \alpha \in \mathrm{K}^k\right\}, \text{ for } k=nirs,\ ndrs,\ crs \tag{2}$$

where $\mathrm{K}^{nirs} = [0,1)$, $\mathrm{K}^{ndrs} = [1,\infty)$ and $\mathrm{K}^{crs} = [0,\infty)$. A technology that exhibits *ndrs*, *nirs* or *crs* in different regions of the production frontier is said to be characterized by variable returns to scale (*vrs*). This production possibility set is denoted Ψ^{vrs}.

It is possible to define an input oriented technical efficiency measure in the sense of Farrell (1957) with respect to the various assumptions concerning the returns to scale. That is:

$$\theta^k(x,y) = \min\left\{\theta \middle| (\theta x, y) \in \Psi^k\right\}, \text{ where } k=\ nirs,\ ndrs,\ crs \text{ and } vrs. \tag{3}$$

From Färe and Grosskopf (1985), the (technical) efficiency measures defined in (3) can be used to construct scale efficiency measures for each organization. A scale efficiency measure is the ratio of the efficiency measure under *crs* technology and a *vrs* type technology. That is:

$$S_1(x,y) = \theta^{crs}(x,y) / \theta^{vrs}(x,y) \le 1. \tag{4}$$

We say that the production technology is of the *crs* type if $S_1(x,y) = 1$. To determine the returns to scale of the technology when $S_1(x,y) < 1$ we compute a second ratio that is less restrictive than the initial ratio. That is:

$$S_2(x,y) = \theta^{nirs}(x,y) / \theta^{vrs}(x,y) \le 1. \tag{5}$$

When $S_2(x,y) = 1$ the technology is said to exhibit decreasing returns to scale and for values strictly less than one, i.e. $S_2(x,y) < 1$, it is said to exhibit increasing returns to scale.

3. Efficiency estimation and statistical model

Efficiency measures are unknown, to estimate them we consider a sample of *n* observations, $\Psi_0 = \left\{(x_j, y_j)\right\}_{j=1}^{n}$. The smallest convex envelop of the sample gives the DEA estimator of Ψ in the *vrs* case. That is:

$$\hat{\Psi}^{vrs} = \left\{(x,y) \in \mathbb{R}^{m+s} \middle| x \ge \sum_{j=1}^{j=n} \lambda_j x_j,\ y \le \sum_{j=1}^{j=n} \lambda_j y_j, \sum_{j=1}^{j=n} \lambda_j = 1 \ \forall\, j = 1,\ldots,n\right\}. \tag{6}$$

To obtain the estimators for the other types of returns to scale, $\hat{\Psi}^{nirs}$ and $\hat{\Psi}^{crs}$ it is sufficient to slightly alter the constraint on the sum of the λ_j. That is, in the *nirs* case we have $\sum_{j=1}^{j=n} \lambda_j \le 1$, and to obtain a *crs* envelop, no constraints on the λ_j are necessary other than non negativity of the λs.

The estimator of Farrell's efficiency is obtained by substituting $\hat{\Psi}^k$ to Ψ^k in equations (3):

$$\hat{\theta}^k(x,y) = \min\left\{\theta \middle| (\theta x, y) \in \hat{\Psi}^k\right\} \quad k=vrs,\ nirs,\ crs. \tag{7}$$

The resulting scale ratio estimators are $\hat{S}_1 = \hat{\theta}^{crs}(x,y) / \hat{\theta}^{vrs}(x,y)$ and $\hat{S}_2 = \hat{\theta}^{nirs}(x,y) / \hat{\theta}^{vrs}(x,y)$.

To insure the consistency of the estimator, it is necessary to specify a statistical model that allows us for a full to characterization of the Data Generating Process (DGP). This is the content of the following assumption.

Assumption A1: The set of observations $\left\{(x_j, y_j)\right\}_{j=1}^{n}$ are identically and independently distributed (*i.i.d.*) random variables with probability density function $f(x,y)$ defined on Ψ.

Assumption A2: The probability density function $f(x,y)$ is continuous on the interior of Ψ, and $f\left(x^{\partial}(y),y\right)>0$ where $x^{\partial}(y)=\theta(x,y)x$ is any point on the frontier of Ψ.

Assumption A3: The efficiency measure $\theta(x,y)$ is differentiable in *x* and *y*.

Assumption A1, A2 and A3 together define the statistical model that allows us to characterize the DGP, denoted $\Im$. In fact, the DGP is entirely characterized by the production possibility set Ψ and the density *f*. That is, $\Im=(\Psi,f)$.

4. Bootstrapped test statistics

The test procedure is in two steps. The first step consists in testing the null that a given combinations (x,y) is scale efficient, i.e. its technology is of the *crs* type. The alternative hypothesis has to be less restrictive. One natural hypothesis is that the combination (x,y) is characterized by a *vrs*-technology. Then we have:

$$Test\#1: H_0: S_1(x,y)=1$$
$$H_A: S_1(x,y)<1$$

If the hypothesis H_0 is rejected, we still have to identify whether the returns to scale are increasing or decreasing. This will work if we can find a "new" null hypothesis that is less restrictive than the one in the first test. One way of doing this is to suppose that under the null hypothesis the combination (x,y) is subject to decreasing returns to scale (*drs*). Then, the alternative hypothesis would be that the combination (x,y) is subject to an increasing return to scale (*irs*) technology. Thus, the second test is:

$$Test\#2: H_0^{'}: S_2(x,y)=1$$
$$H_A^{'}: S_2(x,y)<1$$

The test statistics for the first and second test are $\hat{S}_1=\hat{\theta}^{crs}(x,y)/\hat{\theta}^{vrs}(x,y)$ and $\hat{S}_2=\hat{\theta}^{nirs}(x,y)/\hat{\theta}^{vrs}(x,y)$, respectively. To apply these tests we need to find an approximation of the sampling distribution of the estimators of both scale efficiency ratios, $\hat{S}_1$ and $\hat{S}_2$. This approximation rests on the bootstrap method that consists in identically replicating the empirical DGP many times and study the behavior this set of bootstrapped estimates. To implement the procedure, we first generate, from the original sample Ψ_0, *B* pseudo-samples: Ψ_b^*, $b=1,\ldots,B$. Then, the original estimation method (DEA in our case) is applied to each pseudo-samples to obtain the bootstrap estimator of the test statistic $\hat{S}_1^*$ for $\hat{S}_1$ (and $\hat{S}_2^*$ for $\hat{S}_2$). The later are expressed in terms of $\hat{\theta}^{*crs}$, $\hat{\theta}^{*vrs}$ and $\hat{\theta}^{*nirs}$. To generate the pseudo-efficiencies $\hat{\theta}^{*k}, k=crs,vrs,nirs$, we use an homogenous bootstrap methodology developed by Simar and Wilson (1998). This procedure rests on the assumption that the efficiency structure is homogenous.[2] That is, the efficiency score θ is independent of (η,y): $f(\theta|\eta,y)=f(\theta)$. A consistent estimator of *f*, obtained using a kernel estimator and corrected by Silverman's (1986) is defined as follow:

[2] Because Farrell's measure is radial, we are allowed to write the input vector x in polar coordinates. That is, the modulus of x is $\omega(x)=\|x\|=\sqrt{x^Tx}$ and the angle is $\eta=\eta(x)\in[0,\pi/2]^{m-1}$. This allows us to write the density as $f(x,y)=f(\omega,\eta,y)$.

$$\hat{f}^c(t)=\begin{cases}2\hat{g}(t) & \text{if } t\le 1\\ 0 & \text{otherwise}\end{cases}, \text{ where } \hat{g}(t)=\frac{1}{2nh}\sum_{j=1}^{j=n}\left[\phi\left(\frac{t-\hat{\theta}_j}{h}\right)+\phi\left(\frac{t-2+\hat{\theta}_j}{h}\right)\right]. \quad (8)$$

We use a normal Gaussian kernel, denoted ϕ, and the bandwidth, *h*, is set following the normal reference rule (Silverman (1986)). The pseudo-scores $\hat{\theta}^{*k}, k=crs,vrs,nirs$ are generated from $\hat{f}^c$, in five steps:[3]

Step 1: Compute $\hat{\theta}_j^{crs}=\hat{\theta}_j^{crs}(x_j,y_j)\ \forall\, j=1,\ldots,n$ based on (7).

Step 2: Generate smoothed resampled pseudo-efficiencies as follows. First generate $\{\rho_j^*, j=1,\ldots,n\}$ by resampling with replacement a sample of size *n*, from the empirical distribution $\{\hat{\theta}_j^{crs}, j=1,\ldots,n\}$. Then generate the sequence $\{\tilde{\rho}_j^*, j=1,\ldots,n\}$ as follows:

$$\tilde{\rho}_j^*=\begin{cases}\rho_j^*+h\varepsilon_j^* & \text{if } (\rho_j^*+h\varepsilon_j^*)\le 1\\ 2-(\rho_j^*+h\varepsilon_j^*) & \text{otherwise}\end{cases}, \text{ where } \varepsilon_j^*\sim N(0,1).$$

Then, generate the pseudo-efficiencies γ_j^* using $\gamma_j^*=\bar{\rho}^*+(\tilde{\rho}_j^*-\bar{\rho}^*)/\sqrt{1+h^2/\hat{\sigma}_{\hat{\theta}^k}^2}$, $\bar{\rho}^*=(1/n)\sum_{j=1}^n\rho_j^*$

Step 3: Compute the pseudo variable inputs, $x_j^*=(1/\gamma_j^*)\hat{\theta}_j^{crs}x_j$.

Step 4: Compute the bootstrapped efficiency measures $\hat{\theta}_j^{*crs}$ and $\hat{\theta}_j^{*vrs}$ using the pseudo variable inputs based on the following program:

$\hat{\theta}^{*crs}(x,y)=\min\left\{\theta|\theta x\ge\sum_{j=1}^{j=n}\lambda_j x_j^*,\ y\le\sum_{j=1}^{j=n}\lambda_j y_j,\lambda_j\ge 0\right\}$; and

$\hat{\theta}^{*vrs}(x,y)=\min\left\{\theta|\theta x\ge\sum_{j=1}^{j=n}\lambda_j x_j^*,\ y\le\sum_{j=1}^{j=n}\lambda_j y_j,\sum_{j=1}^{j=n}\lambda_j=1,\ \lambda_j\ge 0\right\}$.

Step 5: Repeat steps 2-5 *B* times to obtain *B* efficiency measures $\{\hat{\theta}_{bj}^{*vrs},\hat{\theta}_{bj}^{*crs}, j=1,\ldots,n, b=1,\ldots,B\}$.

If the test leads to rejection of the null hypothesis we pass to the second test.

Repeat steps 1, 2 and 3 for *k=nirs*.

Step 4: Compute the bootstrapped efficiency measures $\hat{\theta}_j^{*nirs}$ and $\hat{\theta}_j^{*vrs}$ using the pseudo variable inputs based on the following program:

$\hat{\theta}^{*nirs}(x,y)=\min\left\{\theta|\theta x\ge\sum_{j=1}^{j=n}\lambda_j x_j^*,\ y\le\sum_{j=1}^{j=n}\lambda_j y_j,\sum_{j=1}^{j=n}\lambda_j\le 1,\ \lambda_j\ge 0\right\}$; and

$\hat{\theta}^{*vrs}(x,y)=\min\left\{\theta|\theta x\ge\sum_{j=1}^{j=n}\lambda_j x_j^*,\ y\le\sum_{j=1}^{j=n}\lambda_j y_j,\sum_{j=1}^{j=n}\lambda_j=1,\ \lambda_j\ge 0\right\}$.

Step 5: Repeat steps 2-5 *B* times to obtain *B* efficiency measures $\{\hat{\theta}_{bj}^{*vrs},\hat{\theta}_{bj}^{*nirs}, j=1,\ldots,n, b=1,\ldots,B\}$.

The simulation results are used to calculate the pseudo-scores $\hat{S}_1^*=\hat{\theta}^{*crs}/\hat{\theta}^{*vrs}$ and $\hat{S}_2^*=\hat{\theta}^{*nirs}/\hat{\theta}^{*vrs}$. These latter allows us to estimate the empirical distribution of $(\hat{S}_1^*-\hat{S}_1)$ (and of course $(\hat{S}_2^*-\hat{S}_2)$) which used to approximate the unknown distribution of the statistic $(\hat{S}_1-S_1)$. Then for a given size α, we reject null hypothesis when $\Pr(\hat{S}_1^*\le\hat{S}_1)\le\alpha$ for the first test and when $\Pr(\hat{S}_2^*\le\hat{S}_2)\le\alpha$ for the second.

[3] Our approach differs from Simar and Wilson (2002). For the latter, efficiency scores are generated independently of the null hypothesis. This could provide pseudo-values that are not admissible (> 1).

5. A Monte Carlo experiments

In this section we conduct a series of Monte Carlo experiments to evaluate the performance of our test procedure.[4] The experiments are conducted under the following general framework. At each Monte Carlo iteration we use two thousand Bootstrap replications, B=2000, for each unit and the experiment is repeated one thousand times, N=1000. In each Monte Carlo experiment, we compute the test statistic for each DMU, using the data generated for the inputs and the outputs. Then we use the bootstrap procedure presented in the previous section to determine the *p-value* for each DMU. For each Monte Carlo experiment, we estimate the real size of each test as the number of times the null hypothesis is rejected divided by the number of experiments (N), given a nominal size (the theoretical α). We conduct a total of eight experiments. They are all with one output (s=1), but we consider also one or two inputs, $(m \in \{1,2\})$, samples of DMUs of size ten and twenty, $n \in \{10,20\}$, and all experiments are ran for test sizes $\alpha = 0.05$ and $\alpha = 0.01$.

We consider the performance of both tests separately. Let us start with an assessment of the performance of the first test. In order to have a tractable problem, easy to understand, we have used a simple data generating process where the number of output is set to one in all experiments and the number of inputs is set to either one or two depending on the experiment. The inputs of DMU j, (x_{ij}), are assumed to be independently and identically uniformly distributed (*iid*) on the interval $[1,9]$. To generate the output under the null hypothesis of constant returns to scale, we proceed as follows. First we generate a sequence of n independently distributed standard normal noises, $N(0,1)$, v_j for j=1,...,n. Then, the noise is used to generate the output for the j^{th} DMU, (y_j), according to the following constant returns to scale Cobb-Douglas production structure:

$$y_j = \prod_{i=1}^{m} x_{ij}^{1/m} e^{-0.1|v_j|}, \; j = 1,\ldots,n \tag{9}$$

Table 1 summarizes the results of the eight Monte-Carlo experiments related to the performance of test #1. A comparison of the real size from the Monte-Carlo experiment and the nominal size (the theoretical size, α), shows that the results are close to the expected size at both one and five percent and for both sample sizes, confirming the validity of our test procedure. The first part of the table presents the results for n=10. On average, the real size is equal to 0.0443 for $(m = s = 1)$, a number fairly close to the nominal size of α=0.05. When we increase the number of inputs to two, $(m = 2)$, the average real size is now equal to 0.0395. That is, the distortion in the size increases, but this can probably be explained by the curse of dimensionality. Similar remarks apply for the size $\alpha = 0.01$. The real size is equal to 0.0142 for the one-input experiment (*m=1*) and equal to 0.0128 for two-input experiment (*m=2*), which are still very close to the nominal size.

When we increase the sample size to twenty $(n = 20)$, the distortions decrease significantly, to almost 0.001 for both nominal sizes and both experiments with the number of inputs set to either one $(m = 1)$ or two $(m = 2)$. This clearly shows that our test procedure is performing better when the sample size increases, as it would be expected.

[INSERT TABLE 1 HERE]

We now tackle the performance of the second test. To do so, we need to generate data under the null hypothesis of non-increasing returns to scale (*nirs*). We consider separately the case of one input and two inputs.

[4] The Monte Carlo experiment uses a SAS program written by the authors.

In the one input case, given that the $\left(x_{ij}\right)$ are generated as uniform *iid* random variables on the interval $[1,9]$, the output of the j^{th} DMU can be generated as follows:

$$y_j = \begin{cases} x_{1j} e^{-0.1|v_j|} & \text{if } x_{1j} \le 5 \\ \left(\frac{5}{2} + \frac{1}{2} x_{1j}\right) e^{-0.1|v_j|} & \text{otherwise} \end{cases} \tag{10}$$

In the two-input case, we split the returns to scale regions not based on input levels, as in the one input case, but based on a reference output level. The procedure is as follows. First, using the randomly generated inputs, we compute the reference output level for the j^{th} DMU, $\bar{y} = x_{1j}^{1/2} x_{2j}^{1/2}$. Then the output of this DMU is generated according to the following formulae:

$$y_j = \begin{cases} x_{1j}^{1/2} x_{2j}^{1/2} e^{-0.1|v_j|} & \text{if } \bar{y} \le 5 \\ \sqrt{5} x_{1j}^{1/4} x_{2j}^{1/4} e^{-0.1|v_j|} & \text{otherwise} \end{cases} \tag{11}$$

To evaluate the performance of the second test, we proceed as we did with test #1, and we conduct the same eight Monte-Carlo experiments (the combinations of s=1, m=1, 2, α=0.01, 0.05, and n=10, 20). The results of these experiments are summarized in Table 2. The results are very similar to those of Table 1. The simulated sizes are very close to the nominal sizes, again confirming the good performance of our test. When the number of DMU is set to ten, n=10, and $m = s = 1$, the average simulated size is equal to 0.047, a distortion of 0.003 for a nominal test size of 0.05. When the sample of DMU is increased to twenty, $n = 20$, the distortion of the size is even smaller, as it is equal to 0.0023. The distortion increases however when the number of inputs increases to 2 from 1 (all other parameters equal). As in the case of the first test, this may be attributed to the curse of dimensionality. The same phenomena are observed when for size $\alpha = 0.01$. For $m = s = 1$ and $n = 10$, the simulated size is 0.0155 and decreases to 0.01395 when the number of DMUs increases to twenty ($n = 20$). Again, these results confirm our initial conclusions that the performance of the test is good and improves when the size of the sample increases, as expected.

[INSERT TABLE 2 HERE]

6. Conclusion

In this article we have developed a procedure to test non parametric statistical hypothesis concerning the scale efficiency of organization. To assess the performance of our test, we perform limited Monte-Carlo experiments. We chose to limit the number of cases under study because of the complexity of the calculations involved (the number of experiments for each test requires to solve $\left(N \times n \times (B+1)\right)$ linear programs). However, the results largely confirm the very good performance of our test.

References

[1] Farrell, M.J. (1957). The Measurement of Productive Efficiency. *J.Roy.Stat.Soc.* A 120, 253-290.

[2] Färe, R. (1988). Fundamentals of Production Theory. *Lecture Notes in Economics and Mathematical System*s, Vol. 311, Springer-Verlag, Berlin Heidelberg.

[3] Färe, R. and Grosskopf, S. (1985). A Nonparametric Cost Approach to Scale Efficiency. *Scandinavian Journal of Economics* 87, 594-604.

[4] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

[5] Simar, L. and Wilson, P.W. (2002). Non-parametric tests of returns to scale. *European Journal of Operational Research* 139, 115-132.

[6] Simar, L. and Wilson, P.W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science* 44, 49-61.

Table 1: Monte-Carlo estimates of size
Test #1 (H_0 : CRS)

DMU	Nominal size			
	$\alpha = 0.05$		$\alpha = 0.01$	
	$m = s = 1$	$m = 2,\ s = 1$	$m = s = 1$	$m = 2,\ s = 1$
<u>n=10</u>				
1	0.051	0.054	0.021	0.020
2	0.035	0.029	0.012	0.003
3	0.038	0.033	0.011	0.017
4	0.039	0.034	0.018	0.012
5	0.050	0.043	0.017	0.014
6	0.049	0.051	0.015	0.017
7	0.042	0.023	0.012	0.005
8	0.040	0.049	0.007	0.015
9	0.057	0.037	0.013	0.012
10	0.042	0.042	0.016	0.013
Mean	0,0443	0,0395	0,0142	0,0128
<u>n=20</u>				
1	0.046	0.050	0.004	0.020
2	0.055	0.052	0.013	0.028
3	0.050	0.050	0.004	0.022
4	0.054	0.054	0.011	0.010
5	0.047	0.048	0.000	0.016
6	0.044	0.051	0.020	0.019
7	0.058	0.055	0.010	0.016
8	0.041	0.053	0.010	0.022
9	0.047	0.045	0.016	0.014
10	0.059	0.056	0.017	0.003
11	0.042	0.052	0.004	0.030
12	0.045	0.037	0.004	0.009
13	0.045	0.053	0.000	0.020
14	0.040	0.054	0.014	0.023
15	0.048	0.052	0.012	0.018
16	0.049	0.053	0.020	0.020
17	0.057	0.045	0.040	0.030
18	0.043	0.051	0.010	0.000
19	0.046	0.057	0.000	0.018
20	0.049	0.052	0.018	0.022
Mean	0,04825	0,051	0,01135	0,018

Table 2: Monte-Carlo estimates of size
Test #2 (H_0 : NIRS)

Nominal size			
$\alpha = 0.05$		$\alpha = 0.01$	
$m = s = 1$	$m = 2,\ s = 1$	$m = s = 1$	$m = 2,\ s = 1$
0.047	0.060	0.017	0.028
0.043	0.036	0.012	0.012
0.050	0.041	0.018	0.019
0.048	0.039	0.018	0.014
0.057	0.039	0.023	0.015
0.053	0.044	0.017	0.017
0.047	0.028	0.015	0.004
0.041	0.047	0.011	0.014
0.045	0.035	0.015	0.007
0.039	0.057	0.009	0.020
0.047	0.0426	0.0155	0.015
0.058	0.057	0.008	0.018
0.050	0.052	0.017	0.008
0.064	0.120	0.014	0.024
0.049	0.031	0.012	0.010
0.057	0.068	0.020	0.013
0.052	0.070	0.011	0.016
0.048	0.059	0.014	0.018
0.046	0.044	0.016	0.005
0.054	0.056	0.014	0.016
0.038	0.050	0.012	0.022
0.090	0.062	0.016	0.020
0.053	0.062	0.023	0.014
0.044	0.069	0.018	0.026
0.046	0.055	0.007	0.021
0.038	0.038	0.016	0.018
0.051	0.057	0.011	0.019
0.044	0.036	0.006	0.012
0.058	0.024	0.018	0.010
0.052	0.030	0.010	0.018
0.054	0.042	0.016	0.009
0.0523	0.0541	0.01395	0.01585

PCA, FA, ICA and LDA algorithms for Data reduction, Discriminent analysis, Classification and Knowledge extraction of complex biological data

Ali Mohammad-Djafari and Ghazaleh Khodabandelou *

Laboratoire des signaux et systmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France

Abstract. In this paper, first we present, in a unifying way, Principal Component Analysis (PCA), Factor Analysis(FA), Independent Component Analysis (ICA) and Linear and Nonlinear Discriminent Analysis (DA) methods and associated practical algorithms which can be used in Data Reduction (DR) and supervised classification of multivariate great dimentional data. Then, we present A Matlab toolbox which gives the possibility to simulate and test these algorithmes and show simulation results which show the performances of these algorithms. In the second part, we describe some biological experiments related to studying the circadien cell cycles of mouses and cancer treatement: In a first step to observe different kind of data (Tempreature, Activity, Hormons, Genes and Proteins) to understand the complex biological and medical effects. The biologists need to visualize, to analyse and to do classifications on these data and finally to extract some knowledge from them and propose biological models describing the studied phenomena. These data are often complex: multivariate, great dimensionality, heterogeneous, with missing data, and observed at different sampling rates. The classical methods of PCA, FA, ICA and LDA can not directly handle these data. In this paper, first we show the performances of these algorithms on real data and then propose prospective new extentions to push further their limits.

Keywords: Dimensionality reduction, Principal Component Analysis, Factor Analysis, Independent Component Analysis, Linear Discriminent Analysis, Bayesian inference, Sources separation, Inverse problems.

1 Introduction

Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA) and Linear Discriminent Analysis (LDA) methods are the main classical methods for analysing high dimensional data [1, 2, 3, 4, 5, 6, 7, 8]. PCA, FA and ICA are mainly used for dimensionality reduction and LDA for supervised classification. Even if these methods are well defined, still there exist different algorithms for practical usage. PCA and FA are the most stable ones

*This work is a part of ERASYSBIO-C5Sys European project "Circadian and cell cycle clock systems in cancer": http://www.erasysbio.net/index.php?index=272

because they use quadratic criteria and L2 norms (second order statistics in statistical interpretation and Gaussian hypothesis in probabilistic interpretation) and so they are very simple to implement. The caracteristics of the results obtained by PCA and FA are well known. For example, we know that the factors are obtained upto a rotation indetermination. ICA is more complex because the criteria used to be optimized are often non quadratic (Kullback-Leibler divergence) and use higher order statistics (HOS) and non Gaussian probabilities. The corresponding algorithms are then more sophisticated. However the common properties of independent components are that they are obtained upto a permutation and scale factor indetermination. LDA can be considered as a particular supervised classification method where we know the number of classes.

In this paper, in a first step, we present simulation tools: to generate different factors with different properties; to generate different data sets with linear or non linear dependencies; to add different kind of errors; to apply different algorithms of PCA, FA, ICA, LDA, ... and to compare the obtained results.

In a second step, we used these methods for real data set obtained by biologistes working on circadian and cell cycle influence on cancer. This work is done in collaboration within the European project ERASYSBIO. The main object of this paper is to analyse these results and show the performances of these tools in real applications.

This paper is organized as follows: first, we present, in a unifying way, some backgrounds on PCA, FA, ICA and LDA methods and algorithms. Then, we describe more the simulation tools and show the results on simulated data, and finally, we show the results on some real data. The main objective of this paper is to explore new methods which can handle the real data.

2 A unifying presentation of PCA, FA, ICA and LDA

PCA, FA, ICA and LDA are classical methods of dimensionality reduction and data analysis. Due to the origin of these methods, there have been many different presentations and interpretations. Here, we try to present them in an unifying context of forward modeling and inversion. To do this, we start by defining the factors $\boldsymbol{f}(t) = [f_1(t), \cdots, f_N(t)]$ which is an N-dimensional vector of time series. Here, we choosed time series due to our final application.

In a first step, we assume that the observed data $\boldsymbol{g}(t) = [g_1(t), \cdots, g_M(t)]$ are obtained via a mixing (or loading) matrix $\boldsymbol{A}$ of dimensions $[M \times N]$ through the forward model

$$\boldsymbol{g}(t) = \boldsymbol{A}\,\boldsymbol{f}(t) + \boldsymbol{\epsilon}(t), \quad t = 1, \cdots, T \tag{1}$$

where $\boldsymbol{\epsilon}$ represents the errors of modeling and T is the total number of observed samples.

Using this forward model, the objective of many data analysis methods such as PCA, FA, ICA and LDA is to obtain the factor $\boldsymbol{f}$ and the loading matrix $\boldsymbol{A}$. Described as such, we see that this estimation problem is very ill-posed in the sense that we can find many combinations of factors and loading matrix

which can satisfy this model. In the following, we use this model to explain the differences between PCA, FA, ICA and LDA.

2.1 Principal Component and Factor Analysis (PCA and FA)

PCA and FA methods try to find uncorrelated factors $\widehat{\boldsymbol{f}}$. Because correlation describes a linear dependence, the main assumption is then that $\widehat{\boldsymbol{f}}$ has to be obtained through a linear combination of the data: $\widehat{\boldsymbol{f}}(t) = \boldsymbol{B}\,\boldsymbol{g}(t)$, where the matrix $\boldsymbol{B}$ is called separating (or demixing or deloading) matrix.

To summarize PCA and FA algorithms, we can start by writing the expression of the covariance of the data: $\boldsymbol{\Sigma}_g = \text{cov}[\boldsymbol{g}] = \boldsymbol{A}\,\text{cov}[\boldsymbol{f}]\,\boldsymbol{A}' + \text{cov}[\boldsymbol{\epsilon}]$, where it is assumed that $\boldsymbol{\epsilon}$ and $\boldsymbol{f}$ are independent. This relation is the main basis of the two methods PCA and FA. In both methods, it is assumed that the factors are uncorrelated which means that $\text{cov}[\boldsymbol{f}]$ is a diagonal matrix: In PCA, $\text{cov}[\boldsymbol{f}] = \boldsymbol{I}$ and in FA $\text{cov}[\boldsymbol{f}] = \text{diag}\left[\sigma^2_{f_1}, \cdots, \sigma^2_{f_N}\right]$.

The first step in both methods then is to estimate the covariance of the data. A very simple nonparametric method to estimate it is:

$$[\widehat{\boldsymbol{\Sigma}}_g]_{ij} = \frac{1}{T}\sum_{t=1}^{T} g_i(t)\,g_j(t), \quad \forall i,j \in [1,\cdots,M] \tag{2}$$

The basic PCA and FA algorithm is then to decompose this matrix using Singular Value Decomposition (SVD): $\widehat{\boldsymbol{\Sigma}}_g = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ and identify $\boldsymbol{A}$ by $\boldsymbol{U}$ and $\text{cov}[\boldsymbol{f}]$ by $\boldsymbol{\Lambda}$ in the case where the error term $\boldsymbol{\epsilon}$ can be neglected.

The main difference between FA and PCA is just the way to identifying the the matrices $\boldsymbol{A}$ (or $\boldsymbol{B}$) and the factors. For more details, see [1]. At this step, it is important to note that if $\widehat{\boldsymbol{f}}$ is a PCA or FA solution, then, $\boldsymbol{R}\widehat{\boldsymbol{f}}$ where $\boldsymbol{R}$ is a rotation matrix is also a solution. We may say that PCA and FA solution is given upto a rotation matrix. To choose a specific solution, we need to impose more constraints.

2.2 Independent Component and Factor Analysis (ICA)

In PCA the main assumption is that the factors are noncorrelated. In ICA, the main assumption is independence which is much powerfull than uncorrelation. This independence can be imposed in a strict way $p(\boldsymbol{f}) = \prod_{j=1}^{N} p(f_j)$, or by minimizing the Kullback-Leibler (KL) divergence $\text{KL} = \left(\prod_{j=1}^{N} p(f_j) \;:\; p(\boldsymbol{f})\right)$, where $\text{KL}(p:q) = \int p \log p/q$ There are many variants of ICA algorithmes, but all can be considered as particular approximations of these criteria with different optimization algorithms. Here, we do not go much more in details of these algorithms, but it is important to note that ICA solutions have two main indeterminations: scale and permutation.

2.3 Linear Discriminent Analysis (LDA)

In PCA and FA, we were looking for uncorrelated factors, in ICA for independent factors. In these methods, the data were assumed to be homogeneous (no class information). In LDA, we know that the data are classified and we know the number of classes. So, the main objective here is to find the factors such that, when the data are projected on them, they can be separated through $N-1$ hyperplanes. Another way to present the things is to say that, in PCA and FA the assumption is that the sources $\boldsymbol{f}$ are Gaussian, in ICA they are assumed to be Non Gaussian, in LDA they are assumed to be a mixture of Gaussians where the hidden variable of the mixture is the classes. This interpretation gives us the possibility to extend these methods easily.

3 Proposed extensions: Bayesian inference

As we mentionned, looking at the problem as an inference on $\boldsymbol{f}$ and $\boldsymbol{A}$ based on the forward model $\mathcal{M}$ (1), the observed data $\boldsymbol{g}$ and some prior information on the sources and the mixing matrix, the natural tool to use is the Bayesian inference which can be summarized as:
– obtaining an expression for the joint posterior law:

$$p(\boldsymbol{f}, \boldsymbol{A}, \boldsymbol{\theta}|\boldsymbol{g}; \mathcal{M}) = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{A}, \boldsymbol{\theta}_\epsilon; \mathcal{M})\, p(\boldsymbol{f}|\boldsymbol{\theta}_f)\, p(\boldsymbol{A}|\boldsymbol{\theta}_A)\, p(\boldsymbol{\theta})}{p(\boldsymbol{g}|\mathcal{M})} \tag{3}$$

where $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{A}, \boldsymbol{\theta}_\epsilon; \mathcal{M})$ is the likelihood, $p(\boldsymbol{f}|\boldsymbol{\theta}_f)$ and $p(\boldsymbol{A}|\boldsymbol{\theta}_A)$ are the prior on sources and on the mixing matrix and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\epsilon, \boldsymbol{\theta}_f, \boldsymbol{\theta}_A\}$ all the parameters of these laws, and
– using it to infer on all the unknowns of the problem.

Due to lack of place, we do not give more details about this here, but we mention that this approach can give us the possibility to find many classical methods of PCA, FA, LDA, ICA and LDA and much more as particular cases. We can even infer on the number of factors. For more details see [**?**]

4 Presentation of the simulation toolbox

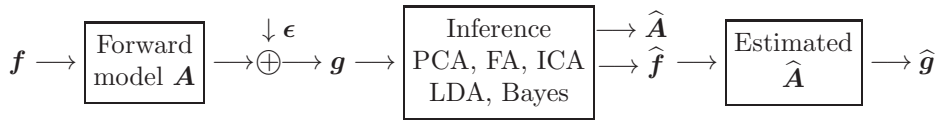
We have developped a menu driven simulation tool, which has, as the main menu, the following steps:
– Generation of different sources (factors) with different properties (Uniform, Gaussian, Mixture of Gaussians, ... ,
– Generation of different data sets with linear or nonlinear dependencies,
– Addition of different kind of errors,
– Application of different algorithms of PCA, FA, ICA, LDA, ... and
– Visualization and evaluation tools which give possibility to evaluate the performances of a given method or to compare the results obtained by two different methods.

As an example of using this simulation tool, we show here a complete set of figures. Figure 1 shows an example of two sources $\boldsymbol{f}$ and five data set $\boldsymbol{g}$ obtained

via a mixing matrix $\boldsymbol{A}$ and addition of some noise $\boldsymbol{\epsilon}$ using the forward model $\boldsymbol{g} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\epsilon}$ and then the results obtained by FA and ICA.

As a second example, we show in Figure 2 two sources generated via a mixture of two Gaussians model. We then again used these sources to generate the data and applied different methods of PCA, FA, ICA (without using the class information) and LDA with using the class information.

As tools to measure the performances of these methods, we propose the following scheme:



and then compare $\widehat{\boldsymbol{g}}$ with $\boldsymbol{g}$, $\widehat{\boldsymbol{f}}$ with $\boldsymbol{f}$, $\widehat{\boldsymbol{A}}$ with $\boldsymbol{A}$, ...

5 Application on real data

As we mentionned, we developped these tools for analysing some biological data in relation with circadian cell cycle and evolution of cancer tumors in the context of the European project ERASYSBIO. A great number of experimentations have been done on mouses. As an example, different quantities such as Temperature, Activity, different Hormones, different Gene expressions and different Proteins are measured during one or a few days and one of the problems addressed is finding the principal components or factors of some of these data. In Figure 2, we show an example of such analysis on Gene expressions time series. For now, we just applied these methods directly of the time series data without accounting for time structure which is very important. However, the results obtained seem to have some significant importance for biologists.

The main difficulties in these data are: great dimensionality (more than fifty), non-homogeneity (Temperature, Activity, Hormones, Genes, Proteins), abberent data, data missing and lack of synchronisation (for example, temperature is measured every 15 minutes but Genes expressions every 3 hours). We need to adapt these methods to account for all these difficulties.

6 Conclusions

In this paper, first we gave a unifying presentation of PCA, FA and ICA based on forward modeling and inversion. This unifying presentation facilitates the comprehension of these different methods. We then presented the main algorithms classically used for these methods. In the second part, we presented a simulation tool which has the possibilities of generating sources and observations, doing FA, PCA and ICA and evaluating the performances of the proposed methods. Finally, we used these tools for analysing some biological data which seems giving important information, or at least confirme their intuition on the role of different quantities.

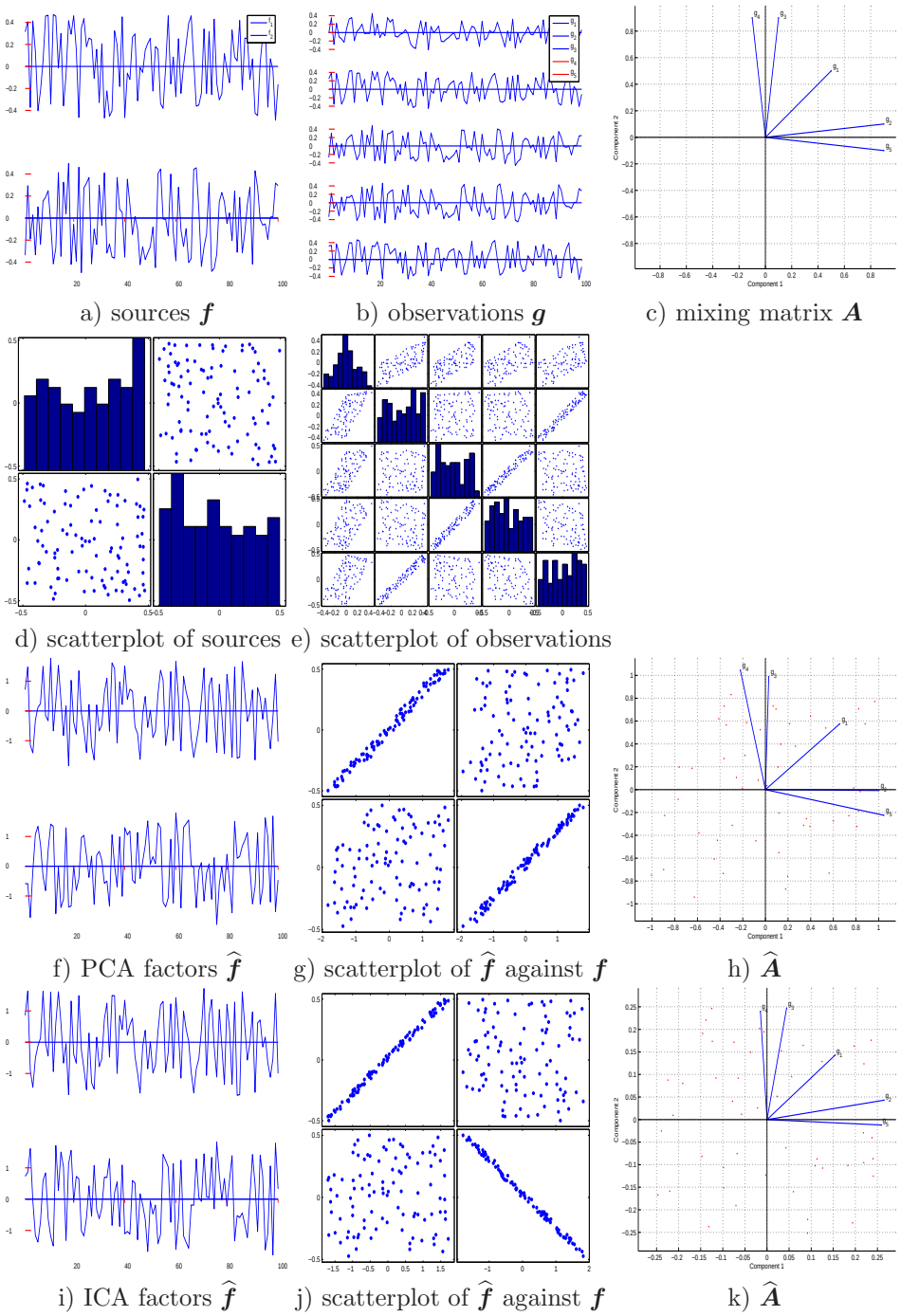
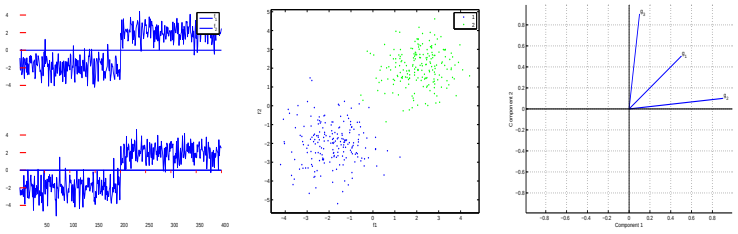
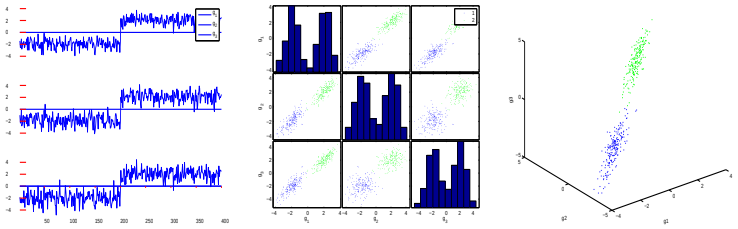


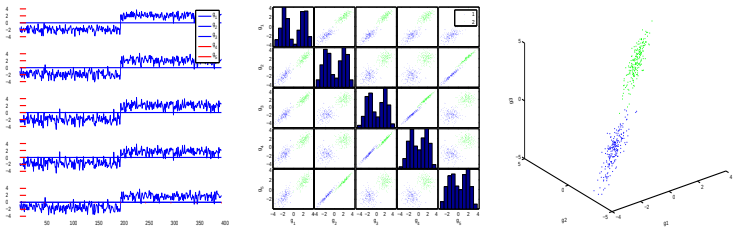
Fig. 1: Simulation of 2 sources $\boldsymbol{f}$ and 5 observations $\boldsymbol{g}$ with $T = 100$ samples: a) sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatterplots of the sources, e) scatterplots of the observations, f) PCA factors $\widehat{\boldsymbol{f}}$, g) scatterplot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$, h) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$, i) ICA factors $\widehat{\boldsymbol{f}}$, j) scatterplot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$, k) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$.



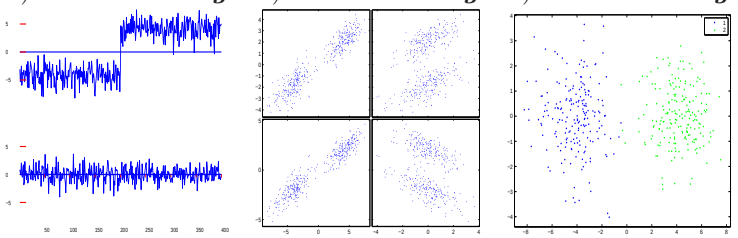
a) 2 sources $\boldsymbol{f}$ b) 2 sources $\boldsymbol{f}$ c) mixing matrix $\boldsymbol{A}$



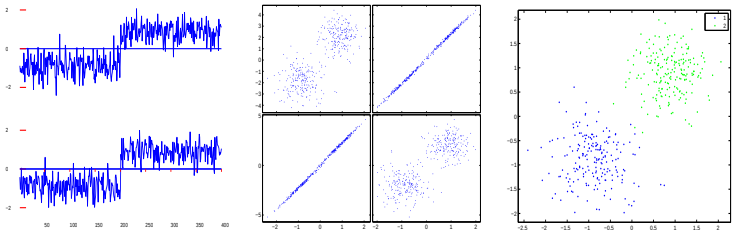
a) 3 observations $\boldsymbol{g}$ b) 3 observations $\boldsymbol{g}$ c) 3 observations $\boldsymbol{g}$



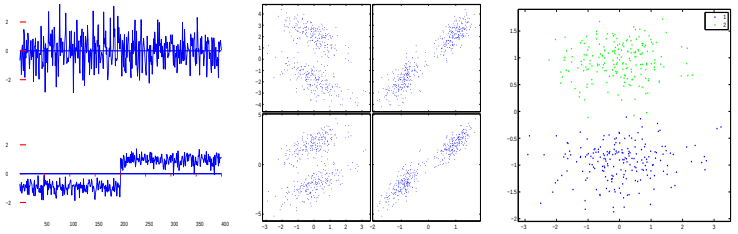
a) 5 observations $\boldsymbol{g}$ b) 5 observations $\boldsymbol{g}$ c) 5 observations $\boldsymbol{g}$



a) PCA $\widehat{\boldsymbol{f}}$ b) PCA $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) PCA $\widehat{\boldsymbol{f}}$

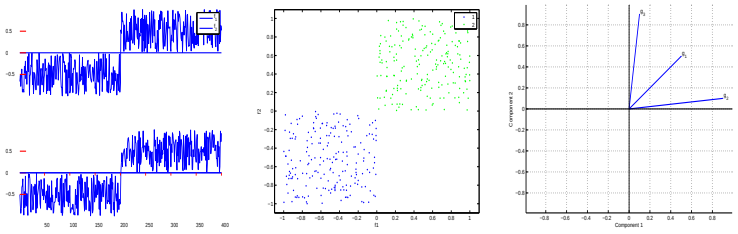


a) factoran $\widehat{\boldsymbol{f}}$ b) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) factoran $\widehat{\boldsymbol{f}}$

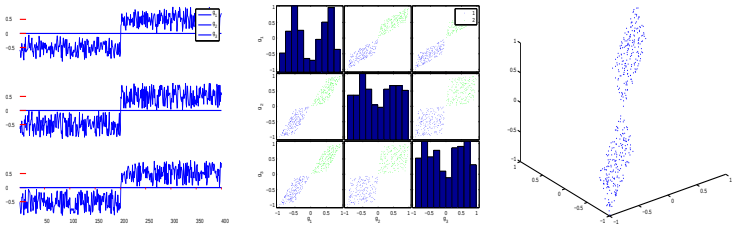


a) factoran $\widehat{\boldsymbol{f}}$ b) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) factoran $\widehat{\boldsymbol{f}}$

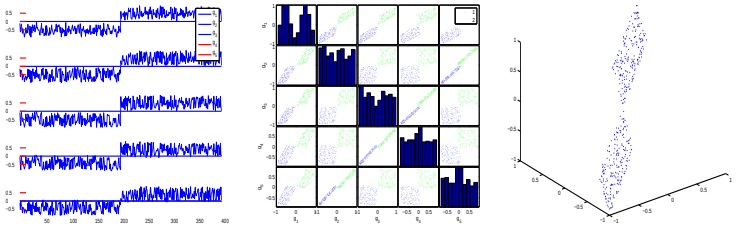
Fig. 2: Simulation of 2 sources $\boldsymbol{f}$ and 3 observations $\boldsymbol{g}$ with $T = 400$ samples: a)sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatterplots of the sources, e) scatterplots of the observations, f) PCA factors $\widehat{\boldsymbol{f}}$, g) scatterplot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ and h) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$.



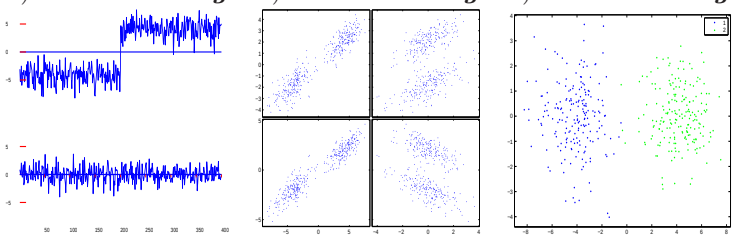
a) 2 sources $\boldsymbol{f}$ b) 2 sources $\boldsymbol{f}$ c) mixing matrix $\boldsymbol{A}$



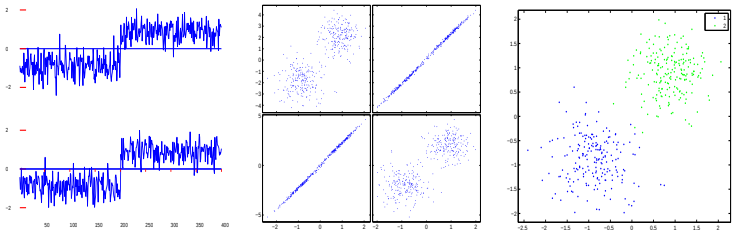
a) 3 observations $\boldsymbol{g}$ b) 3 observations $\boldsymbol{g}$ c) 3 observations $\boldsymbol{g}$



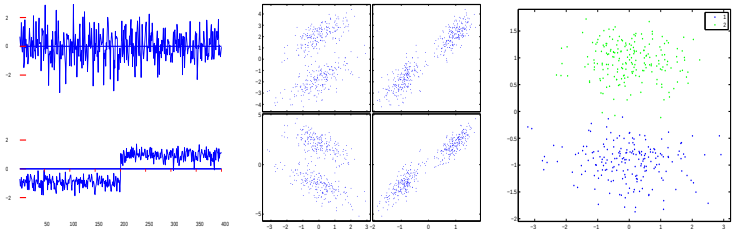
a) 5 observations $\boldsymbol{g}$ b) 5 observations $\boldsymbol{g}$ c) 5 observations $\boldsymbol{g}$



a) PCA $\widehat{\boldsymbol{f}}$ b) PCA $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) PCA $\widehat{\boldsymbol{f}}$

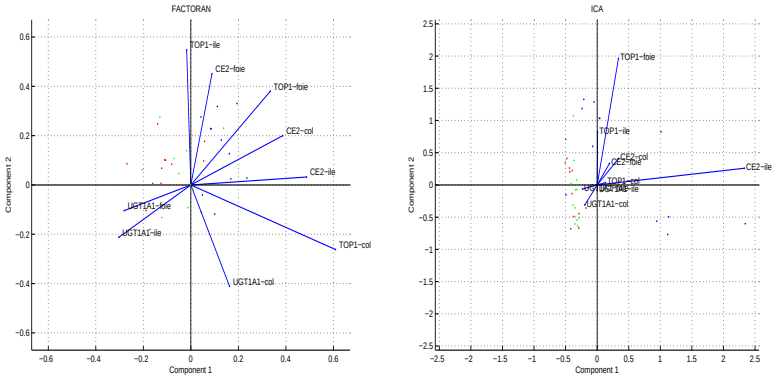


a) factoran $\widehat{\boldsymbol{f}}$ b) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) factoran $\widehat{\boldsymbol{f}}$



a) factoran $\widehat{\boldsymbol{f}}$ b) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ c) factoran $\widehat{\boldsymbol{f}}$

Fig. 3: Simulation of 2 sources $\boldsymbol{f}$ and 3 observations $\boldsymbol{g}$ with $T = 400$ samples: a)sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatterplots of the sources, e) scatterplots of the observations, f) PCA factors $\widehat{\boldsymbol{f}}$, g) scatterplot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ and h) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$.



a) Factor Analysis b) Independent Component Analysis

Fig. 4: A comparison of FA and ICA on the gene expression data

References

[1] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

[2] Pierre Comon. Independent Component Analysis, a new concept ? *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, 36 (3):287–314, April 1994.

[3] Daniel B. Rowe. *Correlated Bayesian Factor analysis*. Phd thesis, Department of Statistics, University of California, Riverside, 1998.

[4] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Technical report, University of Cambridge, Cavindish Laboratory, Cambridge, UK, 1996.

[5] K. Knuth. Bayesian source separation and localization. In A. Mohammad-Djafari, editor, *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, San Diego, CA*, pages 147–158, July 1998.

[6] S. J. Roberts. Independent component analysis: Source assessment, and separation, a Bayesian approach. *IEE Proceedings - Vision, Image, and Signal Processing*, 145(3), 1998.

[7] A. Hyvarinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[8] Michael D. Farrell and Russell M. Mersereau. On the Impact of PCA Dimention Reduction for Hyperspectral Detection of Diffuclt Targets. *Geoscience and Remote Sensing Letters*, 2(2):192–195, 2005.

Improved Dynamic Weighted Majority algorithm for parameter selection

Dhouha Mejri & Mohamed Limam

Larodec, ISG Tunis, 41 Rue de la Libertè Bouchoucha - 2000 Bardo, University of Tunis.

Abstract

Dynamic weighted majority-Winnow (DWM-WIN) algorithm is a powerful classification method that handles nonstationary environment and copes with concept drifting data streams. Although its good performance, this method has a serious problem in choosing the best values of the three parameters β, γ and θ that affects the classification performance of DWM-WIN. Hence, there is a need for a rational automatic selection of parameter values. To deal with this issue, a genetic algorithm (GA) is used as an optimization method to find the best values of these parameters. We have used DWM-WIN as a fitness function of GA. In order to assess this optimized DWM-WIN algorithm, four data sets are simulated from UCI data sets repository to highlight the effectiveness of the optimized DWM-WIN compared to other algorithms.

Key words: Learning and classification, data analysis- data mining.

Résumé

DWM-WIN est un algorithme de classification puissant qui traite l'environnement non stationnaire et s'adapte avec les changements des concepts dans les flux des données. Malgré sa performance, cette méthode possède des problèmes dans le choix des meilleures valeurs de ses 3 paramètres β, γ et θ qui affectent la performance de la classification du DWM-WIN. Alors, une sélection rationnelle automatique des valeurs de ces paramètres est nécessaire. Pour faire face à ce problème, un algorithme génétique (GA) est utilisé comme une méthode d'optimisation afin de trouver les valeurs de ces paramètres. DWM-WIN est utilisé comme une fonction fitness du GA. Afin d'évaluer le DWM-WIN optimisé, quatre base de données ont été simulées à partir de l'UCI dataset pour mettre en évidence l'efficacité de l'optimisation du DWM-WIN par rapport aux autres algorithmes.

Mots-clés: Apprentissage et classification, Analyse des données- Data mining.

1 Introduction

In recent years, ensemble method techniques have attracted the intention of many researchers in classification domain. In almost of the time, the data is arriving in batches and the distribution is nonstationary due to the presence of one or several concept drift in the stream of data. Dynamic weighted majority-winnow (DWM-WIN) algorithm of Mejri et al. (2010) is an advanced ensemble method technique that copes with concept drift. This method improve dynamic weighted majority (DWM) algorithm of Kolter and Maloof (2007) by taking into account the age of the classifiers in the pool as well as the performance of the good classifiers in the ensemble as a new criterion to select the best classifiers in the pool.

Many parameters affect the accuracy rate of the DWM algorithm, β: a parameter for reducing classifier's weight when a classifier makes a wrong prediction and θ: a threshold to remove experts from the ensemble if their weights are under this parameter. Indeed of this two parameters, the improved DWM algorithm noted DWM-WIN uses another parameter γ to increase the weights of good classifiers in the ensemble taking into account their contribution to the learning process.

A simulation study was presented in DWM-WIN of Mejri et al. (2010) to select the value of some parameters like β based on the global accuracy of the algorithm. Other parameters like θ and γ was chosen randomly. Accordingly, this random choice is not rational since it depends on the user's choice which influences the performance of the classification. Indeed, selecting the best parameters can be expensive in terms of time consuming and problems of computations. In this paper, we propose a framework based on GA for parameter selection values that optimizes DWM-WIN algorithm parameters. A set of classifiers aim to determine the fitness function of GA which requires a set of random parameters as initial population.

Many optimization techniques was proposed in the literature. In fact, Stefan et al. (2006) have proposed a Genetic Algorithms (GA) for Support Vector Machine to detect values of kernel parameters. Also, Feng et al. (2008) have applied a multiple feature selection criteria to find the ensemble of features that outperform the classification accuracy.

The paper is outlined as follows: In the second section, an overview of DWM-WIN is presented. Then, the problem of parameter selection is outlined in Section 3. In the fourth section, the application of GA as a solution to optimize DWM-WIN parameters is described. Finally, experimental results are discussed in section 4.

2 Overview of Dynamic weighted majority-Winnow Algorithm

Mining data streams with the presence of concept drift is a very important topic in many applications. In fact, using data chunk is the most adequate technique for dealing with

drifting concept. Furthermore, it does not only allows to determine where nonstationarity exists, but also it helps acquiring knowledge of the old concept used to know the new one.

In DWM-WIN, Mejri et al. (2010) present an advanced ensemble method designed for concept drift based on DWM algorithm of Kolter and Maloof (2007). This improved ensemble method technique was applied in an incrementally updated data, where the underlying data distribution shifts in time according to the target class change. It creates and removes experts based on their performance while considering their age. Subsequently, it copes with experts depending on their weight adjusted during the training phase. Particularly, it reduces the weight of misclassified learners and rewards those contributing to the correct global prediction. Hence, for each new batch, DWM-WIN algorithm dynamically allows finding the best pool of classifiers that give the best performance accuracy and the better classification results when the data is drawn from a nonstationary environment. Although its performance, the improved DWM algorithm suffers from an issue. In the following section we discuss it.

3 Problem of parameter selection

Mejri et al. (2010) have done many cross validations on 5 datasets from the UCI Repository to select the best value of the three parameters: β, γ and θ. Unfortunately, selecting the best parameters can be expensive in terms of time consuming and problems of computations. Hence, the process of choosing parameters value have to be automated with a rational manner. In order to optimize the process of parameter selection, a GA is used to select the best parameter values. The next section describes the use of GA optimizer parameters in DWM-WIN algorithm to find the best values of β, θ and γ.

4 Proposed method: Genetic algorithm for DWM-WIN parameters optimization

4.1 Overview of Genetic Algorithm

GAs was developed by Prof. John Holland and his students at the University of Michigan during the 1960s and 1970s. It is considered as one of the most attractive method of optimization. In fact, Feng et al. (2008) have applied a multiple feature selection criterion that aim to find the ensemble of features having better classification performance than individual algorithms. They used an initialized ensemble of features, then GA searches the optimal subset features from the pool.

In the same context, Belferes and Guitouni (2008) have applied a multivariate genetic algorithm (MGA) optimization method to generate different course of action (COA). The

goal is to determine the largest number of efficient solutions that allow the decision makers (DM) to choose the most appropriate parameters. Stefan et al. (2006) present a model selection technique based on GA that searches the best classifier by re-combining and mutating several number of good classifiers initially constructed in order to ameliorate the classification accuracy.
In the next section, we present details of our approach based on Genetic algorithm as an optimization technique for DWM-WIN algorithm.

4.2 Our approach

Our goal is to find a population of best weights for every parameters values that minimizes the classification error rate. The idea of optimizing DWM-WIN parameters is to consider the accuracy rate of the ensemble of classifiers dynamically trained as a fitness function in GA that searches the best solutions from the initialized population. Our goal is to propose an optimized classification method that automate the choose of its parameter values when dealing with concept drifting data streams.

In fact, GA finds the optimal combination of parameters from the initial population. It searches an ensemble of hypothesis that consists on different combination of the three parameters θ, γ and β called initial population. Each hypothesis is evaluated according to the fitness function of the last population. The used function is represented by a vector $(m, 1)$ where m is the number of initial combination. Each value of this vector represents the accuracy of DWM-WIN algorithm for each hypothesis. After the three steps of GA: selection, crossover and mutation, a new population is created updating the parameter selection process each time a new generation is presented. In the followings, we give a description of different steps of the optimized ensemble method technique.

4.2.1 Initialized Population

A collection of subsets of parameters are selected with GA to find the optimal parameters subsets (β, γ, θ). A population of size (m, n) is constructed, where m is the population size and n is the number of parameters.

4.2.2 Representation of the hypothesis

Each subset of the parameters are encoded with n bits binary vectors. 1 represents the parameter to be selected and 0, the parameter to don't choose.

4.2.3 Fitness function

The fitness function is an objective function that makes use of a population that artificially reproduces test solution. It aims to search the optimal solution from the initialized population. In this paper, an ensemble of classifier is constructed to evaluate the fitness

function of each subsets. This objective function is the accuracy rate of an ensemble of classifiers trained together. Many algorithms can be incorporated into genetic algorithm such as Naive Bays, decision trees, boosting, ect... In this paper, we choose a set of decision trees as base classifiers to construct the optimal DWM-WIN algorithm.

4.2.4 Operators of genetic algorithm

Selection Each individual from the dataset have a proportional value to the fitness function. This proportion is the probability that it will be selected.

Crossover After the selection step, GA randomly chooses a crossover point to divide individuals into two categories: first parents and second ones.

Mutation This step represents the random choose on bits that are reversed in each new generation. The fitness function is represented by the improved classification method. It is computed for each subset of the initialized parameters until finding the best subset of parameters using selection, crossover and mutation as operators.

5 Experiments

Implementations was done in Matlab. DWM-WIN was induced as a fitness function in GA. Four datasets was used for the experiments: Pima, Iris, Tictactoe and German dataset. For DWM-WIN, two numbers of batches are used: 40 and 5. For Genetic algorithm parameters, population size was fixed at 50 and the number of runs used is 10. Experimental results have shown that employing GA as an optimization algorithm for DWM-WIN improve the accuracy rate in 4 datasets from the UCI repository. Table 1 and 2 show a comparison between the average accuracy rate of DWM-WIN before optimization and the improved DWM-WIN using the GA optimization.

Table 1: Average accuracy rates of DWM-WIN and the optimized DWM-WIN with GA using 5 batches.

Datasets	DWMWIN	DWM-WIN using GA
Pima	0.6076	0.791
Iris	0.859	0.886
Australian	0.79	0.776
Tictactoe	0.534	0.53

Table 2: Average accuracy rates of DWM-WIN and the optimized DWM-WIN with GA using 40 batches.

Datasets	DWMWIN	DWMWIN using GA
Pima	0.7681	0.769
Iris	0.6654	0.673
Australian	0.8469	0.859
Tictactoe	0.7907	0.79

Conclusion

We propose an improved DWM-WIN algorithm based on GA as an optimization technique in order improve the classification accuracy. A combination of several classifiers using a dynamic ensemble method technique with GA optimization leads to an improvement of the accuracy rate. This successful optimization technique of a dynamic ensemble method technique is adaptable for different population size and for several number of batches and automates the parameters values for each batch. In this work, we propose a parameter selection optimized method of DWM-WIN. For future work, we plan making a feature selection optimization and applying this improved algorithm for multivariate statistical process control.

Bibliographie

[1] Asuncion, A. and Newman, D. J., (2007), UCI Machine Learning Repository, Department of Information and Computer Sciences, University of California, Irvine, School of Information and Computer Sciences, http://www.ics.uci.edu/mlearn/ MLRepository.html.

[2] Feng T, Xuezheng F, Yanqing Z and Anu G. B., (2008), A genetic algorithm-based method for feature subset selection, Soft Comput, 12:111–120.

[3] Kolter, Z. J. and Maloof, M. A., (2007), Dynamic weighted majority: An ensemble Method for Drifting Concepts *Journal of Machine Learning Research*, 8, 2755–2790.

[4] Mejri, D., Khanchel, R. and Limam, M., (2010), An ensemble method for concept drift in nonstationary environment, *Master thesis, High Institute of management in Tunisia.* .

[5] Stefan, L., Robert, S. and Sven F. C., (2006), Genetic Algorithms for Support Vector Machine Model Selection, International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada.

COMPARAISON DE TAUX D'INCIDENCE PAR DES MODELES DE REGRESSION DERIVES DE POISSON

Sandrine Domecq[a], Marion Kret[a], Christelle Minodier[b], Philippe Michel[a]

a. CCECQA, Hôpital Xavier Arnozan 33604 Pessac Cedex
b. DREES, 14 avenue Duquesne 75 350 Paris 07

Les taux d'incidence des événements indésirables graves observés au cours de deux enquêtes statistiques en milieu hospitalier menées en 2004 et en 2009 ont été comparés. Le nombre d'événements indésirables graves associés aux soins, observés au sein d'unités d'hospitalisation sur 7 jours maximum, a été considéré comme la réalisation d'une variable aléatoire discrète suivant une loi de Poisson [1]. Les taux d'incidence ont été comparés en utilisant des modèles de régression de Poisson ou binomials négatifs en cas de surdispersion [2]. La surdispersion a été vérifiée grâce au test de Dean. Le nombre de jours d'observation a été pris en compte dans un terme offset. Des caractéristiques au niveau des unités d'hospitalisation ont été intégrées dans les modèles comme variables d'ajustement. Des comparaisons ont été faites sur des sous-échantillons d'événements indésirables graves. L'analyse a porté sur 8754 séjours hospitaliers représentant 35234 jours d'observation dans 294 unités d'hospitalisation en 2004 et sur 8269 séjours représentant 31663 jours d'observation dans 251 unités d'hospitalisation en 2009 [3]. L'analyse multivariée a permis d'interpréter les différences entre taux d'incidence en termes de risques relatifs en tenant compte de plusieurs variables d'ajustement et du temps d'exposition. Cette application illustre l'intérêt d'utiliser des modèles de régression dérivés de Poisson par rapport aux méthodes de standardisation pour comparer des taux d'incidence [4].

Mots clés : taux d'incidence, modèle de Poisson, modèle binomial négatif.

The incidence rates of associated adverse events observed during two statistical studies carried out in hospitals in 2004 and 2009 were compared. The number of healthcare associated adverse events, observed in inpatient units over 7 maximum days, was the realization of a discrete random variable from Poisson distribution [1]. The incidence rates were compared using Poisson regression or negative binomial regression to take into account of overdispersion [2]. The Dean's Test was used to test overdispersion. The number of days of observation was included in offset term. Characteristics at the inpatient units were incorporated into the models as covariates. Many comparisons were made on subsamples of healthcare associated adverse events. The analysis included 8754 admissions representing 35,234 days of observation in 294 inpatient units in 2004 and 8269 admissions representing 31,663 days of observation in 251 inpatient units in 2009 [3]. Thus, the differences between incidence rates were interpreted in terms of relative risks, taking into account several covariates and exposure time. This application illustrates the benefits of using Poisson regression compared to the standardization [4].

Keywords: incidence rate, Poisson regression, negative binomial regression

Bibliographie

[1] Allain E, Brenac T. (2001) Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers : le modèle de Poisson et ses extensions. *Recherche Transports Sécurité*, 72, 3-18.
[2] Bouche G, Lepage B, Migeot V, Ingrand P. (2009) Intérêt de la détection et de la prise en compte d'une surdispersion dans un modèle de Poisson : illustration à partir d'un exemple. *RESP*, 57, 285-296.

[3] Michel P, Minodier C, Lathelize M et al. (2010). Les événements indésirables graves associés aux soins observés dans les établissements de santé : résultats des enquêtes nationales menées en 2009 et 2004. *Dossiers Solidarité et Santé DREES*, 10.
[4] Bouyer J, Hémon D, Cordier S et al. (1995) *Epidémiologie Principes et méthodes quantitatives*, Les éditions INSERM, Paris.

Introduction

Deux études nationales, les enquêtes ENEIS (Etudes Nationales sur les Evénements Indésirables graves liés aux Soins), ont été réalisées en 2004 et 2009 pour mesurer et comparer l'incidence des événements indésirables graves associés aux soins (EIG) identifiés pendant une hospitalisation dans les établissements de santé publics et privés en France et d'en connaître la part évitable. Elles ont été financées par la Direction de la Recherche, des Etudes et de l'Evaluation et des Statistiques (DREES) et réalisées par le Comité de Coordination de l'Evaluation Clinique et de la Qualité en Aquitaine (CCECQA).

Objectif

Illustrer l'utilisation des modèles de régression dérivés de Poisson pour comparer des taux d'incidence

Méthode

Il s'agissait d'études d'incidence sur une population de patients hospitalisés et observés dans les établissements de santé publics et privés pendant 7 jours au maximum. Les définitions, le protocole d'enquête, la période de collecte et la maîtrise d'œuvre étaient identiques au cours des deux études. Les échantillons étaient randomisés, stratifiés, avec sondage en grappe à trois degrés (départements, établissements et unités d'hospitalisation de médecine et de chirurgie).
Afin de permettre la meilleure comparabilité possible entre 2004 et 2009, les différences dans les méthodes d'échantillonnage ont été prises en compte lors des traitements statistiques grâce à un redressement des données (repondération pour tenir compte des spécificités du plan d'échantillonnage et de la non-réponse).

Le taux d'incidence des EIG en cours d'hospitalisation a été défini comme le nombre d'EIG observés sur 7 jours au maximum au sein d'unités d'hospitalisation rapporté au nombre total de jours d'hospitalisation observés. Il était exprimé pour 1000 jours d'hospitalisation. Pour la comparaison des taux entre 2004 et 2009, le nombre d'EIG a été considéré comme la réalisation y d'une variable aléatoire discrète Y suivant une loi de Poisson de paramètre μ :

$$\Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad \text{avec } y = 0, 1, 2\ldots$$

et

$$E(Y) = Var(Y) = \mu$$

La comparaison des taux d'incidence a été menée en utilisant des modèles de régression dérivés de Poisson. Les modèles de régression de Poisson sont des modèles linéaires généralisés qui s'appliquent sur des données groupées, en spécifiant la loi de Poisson comme loi de probabilité, la fonction logarithmique comme fonction de lien et une combinaison linéaire comme relation entre les variables explicatives.
Le modèle de Poisson était utilisé lorsque la condition d'égalité entre la moyenne et la variance était vérifiée. Cependant, dans de nombreuses situations le modèle de Poisson est inadéquat car la variance des données est supérieure à la moyenne. Dans ce cas, on se trouve en présence de surdispersion. L'ajustement d'un modèle linéaire généralisé dépend de la fonction de variance et non de la forme précise de la distribution. Ainsi, la surdispersion peut être prise en compte en modifiant la fonction de variance et notamment en utilisant celle du modèle binomial négatif :

$$Var(Y_i) = E(Y_i) + \frac{1}{k}E^2(Y_i) = \mu_i + \frac{1}{k}{\mu_i}^2$$

avec : Y_i : nombre d'EIG dans l'unité d'hospitalisation i
i : indice des unités d'hospitalisation, i = 1, …, I
$\frac{1}{k}$: paramètre de dispersion

La détection d'une surdispersion nécessite d'ajuster le modèle de Poisson simple, puis de réaliser le test de Dean en calculant la statistique T_a telle que :

$$T_a = \frac{\sum_{i=1}^{n}\{(Y_i-\hat{\mu}_i)^2-Y_i+\hat{h}_{ii}\hat{\mu}_i\}}{(2\sum_{i=1}^{n}\hat{\mu}_i^2)^{\frac{1}{2}}} \text{ avec } T_a \sim N(0,1)$$

avec : $\hat{\mu}_i$: valeurs estimées par le modèle de régression de Poisson
$\hat{h}_{ii}$: valeurs levier ou « leverage » des observations (diagonale de la matrice H « hat matrice »)

La variable à expliquer était Y_i le nombre d'EIG comptabilisés au sein de chaque unité d'hospitalisation i. Les quatre variables explicatives X_k introduites systématiquement étaient :

- X_1 : l'année d'étude (2004 vs 2009 - sous forme binaire),
- X_2 : le type d'établissement (Centre Hospitalier Universitaire ou régional (CHU/CHR), centre hospitalier, autre établissement public ou privé à but non lucratif (CH) ou Etablissement Privé (EP) - sous forme d'indicatrices)
- X_3 : le type d'activité (médecine ou chirurgie - sous forme binaire)
- X_4 : l'âge des patients (âge médian des patients de chaque unité d'hospitalisation – sous forme continue)

Le type d'établissement, le type d'activité et l'âge des patients étaient considérés comme des variables d'ajustement. Cela a permis de prendre en compte, d'une part, les particularités des unités d'hospitalisation liées au statut de leur établissement et à leur activité et d'autre part, la différence de l'âge des patients des échantillons mise en évidence entre 2004 et 2009.
Le nombre de jours d'observation a été introduit comme terme offset afin de prendre en compte le temps d'exposition.

Si on exprime le nombre d'EIG dans chaque unité d'hospitalisation en fonction des taux d'incidence, on obtient :

$$Y_i = n_i\lambda_i \quad \text{et} \quad E(Y_i) = E(n_i\lambda_i) = \mu_i$$

avec : n_i : le nombre de jours d'hospitalisation observés dans l'unité d'hospitalisation i
λ_i : le taux d'incidence dans l'unité d'hospitalisation i

L'écriture du modèle est alors :

$$\begin{aligned} \ln[E(Y_i)] &= \ln(n_i\lambda_i) \\ &= \ln(n_i) + \ln(\lambda_i) \\ &= \ln(n_i) + \alpha_i + \beta_k X_{ki} \end{aligned}$$

avec : $\ln(n_i)$: terme offset
α_i : paramètres nuisibles représentant les effets de la variable de stratification (les unités d'hospitalisation)
β_k : vecteur des coefficients de régression, k=1, …,4
X_{ki} : vecteur des variables explicatives

Les paramètres à estimer par la méthode du maximum de vraisemblance étaient les α_i et les β_k.

Des comparaisons des taux d'incidence ont été réalisées pour l'ensemble des EIG, les EIG évitables, les EIG selon les conséquences pour le patient (décès, prolongation du séjour, incapacité, mise en jeu du pronostic vital), et les EIG selon le type d'exposition (procédures invasives, produits de santé (PS), infections).
Une différence des taux d'incidence était interprétée comme significative si le risque relatif (RR) entre 2004 et 2009 était significativement différent de 1. Le RR était obtenu en calculant : exp ($\hat{\beta}_1$), $\hat{\beta}_1$ correspondant au coefficient de régression associé à la variable X_1, l'année d'étude dans le modèle.

Les analyses ont été réalisées sur le logiciel Stata. Les modélisations ont été réalisées grâce aux commandes *poisson* (modèle de Poisson) ou *nbreg* (modèle binomial négatif) avec prise en compte du redressement des données par la commande-préfixe *svy :*.

Résultats

L'analyse a porté sur 8754 séjours hospitaliers représentant 35234 jours d'observation dans 294 unités d'hospitalisation en 2004 et sur 8269 séjours représentant 31663 jours d'observation dans 251 unités d'hospitalisation en 2009. Les taux d'incidence d'EIG figurant dans les tableaux 1 et 2 ont été calculés après redressement des données.

Tableau 1. Comparaison des taux d'incidence d'EIG entre 2004 et 2009 : estimations des risques relatifs

Sur l'ensemble des EIG	2004			2009			RR*	IC à 95%
	nb	(‰)	IC à 95%	nb	(‰)	IC à 95%		
Incidence des EIG	255	(7,2)	[5,7 – 8,6]	214	(6,2)	[5,1 – 7,3]	0,93[bn]	[0,68 - 1,27]
Type de conséquences								
Prolongation	200	(5,7)	[4,4 – 7,1]	153	(4,2)	[3,4 – 5,1]	0,84 [bn]	[0,58 - 1,23]
Pronostic vital	92	(2,5)	[1,7 – 3,3]	71	(2,1)	[1,4 – 2,7]	0,77 [bn]	[0,47 - 1,24]
Incapacité	55	(1,6)	[0,9 – 2,3]	61	(2,0)	[1,4 – 2,7]	1,36 [bn]	[0,83 - 2,23]
Décès	21	(0,5)	[0,2 – 0,9]	16	(0,5)	[0,2 – 0,9]	0,84 [bn]	[0,32 – 2,19]
Type d'exposition								
Procédure invasive	191	(5,5)	[4,2 – 6,9]	159	(4,3)	[3,5 – 5,1]	0,93 [bn]	[0,67 - 1,30]
Produit de santé (PS)	71	(1,6)	[1,2 – 2,1]	83	(2,5)	[1,7 – 3,2]	1,40 [bn]	[0,89 - 2,23]
Infection liée aux soins	57	(1,3)	[0,9 – 1,7]	63	(1,9)	[1,3 – 2,5]	1,53 [bn]	[0,92 - 2,54]
Aucun des 3 types d'expositions	22	(0,6)	[0,3 – 0,9]	12	(0,4)	[0,1 - 0,8]	0,57 [p]	[0,19 - 1,73]

* RR : Risque relatif de 2009 par rapport à 2004 ajusté sur âge médian des patients, spécialité (médecine ou chirurgie) et type d'établissement (CHU-CHR, CH, EP) des unités de soins.
p Utilisation du modèle de Poisson pour estimer le RR en l'absence d'une surdispersion (test de Dean non significatif)
bn Utilisation du modèle Binomial négatif pour estimer le RR en présence d'une surdispersion significative.

Tableau 2. Comparaison des taux d'incidence d'EIG évitables entre 2004 et 2009 : estimations des risques relatifs

Sur les EIG évitables	2004			2009			RR *	IC à 95%
	nb	(‰)	IC à 95%	nb	(‰)	IC à 95%		
Incidence des EIG évitables	95	(2,7)	[1,9 – 3,6]	87	(2,6)	[1,8 - 3,3]	0,98 bn	[0,62 - 1,56]
Type de conséquences								
Prolongation	72	(2,2)	[1,3 – 3,0]	61	(1,7)	[1,1 - 2,3]	0,90 bn	[0,53 - 1,54]
Pronostic vital	39	(1,0)	[0,6 – 1,4]	31	(0,7)	[0,4 - 1,1]	0,72 bn	[0,36 - 1,45]
Incapacité	19	(0,7)	[0,2 – 1,1]	25	(0,7)	[0,4 - 1,0]	1,22 bn	[0,57 - 2,61]
Décès	8	(0,2)	[0,0 – 0,4]	8	(0,4)	[0,0 - 0,7]	1,21 p	[0,31 - 4,66]
Type d'exposition								
Procédure invasive	66	(2,0)	[1,2 – 2,8]	58	(1,7)	[1,1 - 2,3]	1,11 bn	[0,65 - 1,89]
Produit de santé (PS)	30	(0,7)	[0,4 – 1,0]	41	(1,1)	[0,6 - 1,6]	1,39 bn	[0,69 - 2,79]
Infection liée aux soins	17	(0,5)	[0,2 – 0,8]	28	(0,9)	[0,4 - 1,3]	1,79 bn	[0,84 - 3,82]
Aucun des 3 types d'expositions	12	(0,3)	[0,1 – 0,5]	6	(0,1)	[0,0 - 0,3]	0,38 p	[0,11 - 1,31]

* RR : Risque relatif de 2009 par rapport à 2004 ajusté sur âge médian des patients, spécialité (médecine ou chirurgie) et type d'établissement (CHU-CHR, CH, EP) des unités de soins.
p Utilisation du modèle de Poisson pour estimer le RR en l'absence d'une surdispersion (test de Dean non significatif)
bn Utilisation du modèle Binomial négatif pour estimer le RR en présence d'une surdispersion significative.

Dans la majorité des situations, le test de Dean était significatif, justifiant la prise en compte d'une surdispersion dans le modèle de régression. Ainsi, le modèle Binomial négatif a été utilisé à 15 reprises pour estimer les risques relatifs sur 18 comparaisons entre 2004 et 2009 des taux d'incidence d'EIG.
Aucune différence de taux d'incidence des EIG entre 2004 et 2009 n'apparaît comme statistiquement significative.

Conclusion

La méthodologie identique et la reproductibilité des résultats a permis une analyse des évolutions entre 2004 et 2009.
Cette application de l'utilisation des modèles de régression dérivés de Poisson dans le cadre d'une étude épidémiologique illustre leurs avantages sur la standardisation. En effet, la standardisation permet de comparer des taux bruts d'incidence entre deux échantillons différents en éliminant l'effet de structure. Cet effet de structure correspond à des différences de proportions entre les deux échantillons, relatives à une variable donnée, habituellement l'âge. Or, dans de nombreuses situations, il est nécessaire de prendre en compte plusieurs variables explicatives comme cela a été le cas lors de notre analyse. Les modèles de régression constituent une bonne réponse à cette problématique.
De plus, grâce à l'introduction d'un terme offset, les modèles de régression dérivés de Poisson permettent d'étudier la relation de type « dose-effet » entre la survenue d'un événement et le nombre de jours d'exposition.
Par ailleurs, le modèle binomial négatif a permis de prendre en compte la surdispersion dans les données par rapport au modèle de Poisson, fréquemment observée lors de notre analyse. Cela a évité de conclure à tort à l'existence d'une association significative, bien qu'aucune différence n'était significative.
La stabilité des résultats entre 2004 et 2009 est évidemment un résultat en soi mais ne signifie pas absence complète d'évolution entre 2004 et 2009. En effet, la « granularité » de l'enquête ENEIS n'est pas très importante et les indicateurs recouvrent des pans de pratiques et d'organisations très larges. Notamment, l'enquête ENEIS n'est pas en mesure de montrer des résultats de programmes ou actions en gestion des risques sectoriels fins. Elle ne permet donc pas de conclure à l'absence de

changements en termes de culture de sécurité et de comportements des acteurs du système de santé (non mesurés par les indicateurs), ni même à l'absence de résultats des actions en cours : l'augmentation de la complexité technique des actes et des contraintes organisationnelles et budgétaires, avérée sur la période étudiée, aurait notamment pu conduire à une augmentation des risques et de la fréquence des EIG.

Conception de chimiothèques enrichies en inhibiteurs d'interactions protéine-protéine

Christelle Reynès[1], Anne-Claude Camproux[1], Bruno Villoutreix[1] & Olivier Sperandio[1]

[1] *MTi, Unité Inserm U973, Université Paris Diderot, 75013 Paris, FRANCE*

Résumé

Les interactions protéine-protéine (IPP) sont susceptibles de devenir une catégorie majeure de cibles thérapeutiques. Actuellement, une infime partie des molécules thérapeutiques disponibles cible ce type d'interaction, alors que la taille de l'interactome humain semble être très importante (650,000 interactions estimées). Les caractéristiques biologiques de ces systèmes empêchent d'utiliser efficacement les filtrages traditionnels pour molécules thérapeutiques. Il y a donc un réel intérêt à mieux comprendre l'espace chimique recouvert par les inhibiteurs d'IPP pour pouvoir concevoir des chimiothèques de composants orientées vers ce type de chimie. Des arbres de classification ont été utilisés pour construire un modèle distinguant les inhibiteurs de PPI d'autres types de molécules thérapeutiques. Le modèle obtenu a mis en évidence deux descripteurs reflétant des formes moléculaires spécifiques ainsi que la présence d'un certain nombre de liaisons aromatiques. Deux versions de ce modèle ont été implémentées dans un programme gratuit (PPI-HitProfiler) permettant de savoir si une molécule est susceptible d'être un inhibiteur de PPI.
Mots-clés: conception de médicaments, arbres de décision, SVM, interactions protéine-protéine.

Abstract

Protein-protein interactions (PPI) are likely to become a major class of therapeutic targets. So far, only a tiny part of drugs target this kind of interaction whereas the size of human interactome seems to be very important (650,000 estimated PPIs). The specific biological characteristics of those systems prevent any efficient use of conventional screening methods for drugs. Hence, it could be very interesting to have a better understanding of PPI inhibitor chemical space in order to be able to design chemical libraries focused towards this kind of compounds. Classification trees have been used to build a model allowing to discriminate between PPI inhibitors and other kinds of drugs. The obtained model exhibited two molecular descriptors characterizing specific molecular shapes and the presence of a privileged number of aromatic bonds. Two variants of the model have been implemented in a free computer program (PPI-HitProfiler) allowing any user to submit a molecule and to know if it is likely to be a PPI inhibitor.
Keywords: drug design, decision trees, SVM, protein-protein interactions.

1 Introduction

Les interactions entre protéines sont à la base de nombreux aspects du fonctionnement des organismes (Stumpf et al., 2008). Par voie de conséquence, leurs disfonctionnements sont la cause de très nombreuses maladies. Actuellement, ces interactions ne sont la cible que d'une minorité de molécules thérapeutiques inhibitrices et les décideurs hésitent à investir dans ce domaine. En effet, la modélisation des PPI est très complexe, non seulement parce que ces interfaces sont souvent assez flexibles mais également parce que les chimiothèques disponibles sont connues pour ne recouvrir que très partiellement l'espace chimique des inhibiteurs de PPI. Pourtant, les récentes recherches ont permis des avancées thérapeutiques importantes dans ce domaine (Arkin and Whitty, 2009).
Dans ce contexte, une solution pour alléger le temps et le coût de la recherche pourrait être de proposer des chimiothèques enrichies en PPI, comportant des composés plus en phase avec l'espace chimique des inhibiteurs de PPI. La présente étude (Reynès et al., 2010; Sperandio et al., 2010) a eu pour objectif la mise en place d'un modèle permettant de filtrer des composés chimiques et de décider si oui ou non il est judicieux de les inclure dans une étude visant à trouver des inhibiteurs de PPI. Dans un premier temps, les données seront présentées, puis la construction du modèle sera développée, enfin le modèle sera appliqué aux données et interprété.

2 Les données

2.1 Echantillon d'apprentissage

145 inhibiteurs connus de PPI ont été extraits de la littérature. Un filtrage basé sur des critères ADMET classiques assez lâches a permis d'arriver à 81 molécules, puis à 66 molécules après élimination des redondances chimiques. Pour définir un échantillon de molécules non inhibitrices de PPI, un protocole similaire a été utilisé. Le problème était de savoir de quel type de *négatifs* constituer ce jeu. En effet, il est impossible d'être certain qu'une molécule n'inhibe aucune PPI. Afin d'éviter d'introduire trop de faux-négatifs dans le jeu d'apprentissage, nous avons utilisé les composés proposés par une base de données appelée *DrugBank* qui contient 4857 molécules thérapeutiques dont la cible est connue. Comme nous l'avons vu précédemment, les molécules thérapeutiques actuelles incluent très peu d'inhibiteurs de PPI. Cette base de données a donc été utilisée pour constituer les *négatifs* du jeu d'apprentissage. Les mêmes filtres ADMET et de diversité que pour les positifs ont été appliqués. Finalement, 557 molécules ont ainsi été retenues dans la DrugBank. Le déséquilibre d'effectif entre les deux classes est indispensable pour réaliser un apprentissage efficace car dans la réalité, quelle que soit la banque filtrée, on aura beaucoup plus de chances d'avoir un non-inhibiteur de PPI qu'un inhibiteur. Il faut donc que le modèle choisi soit capable de travailler avec des groupes très déséquilibrés.
Concernant les variables, nous avons utilisé les descripteurs moléculaires Dragon (Todes-

chini et al., 2009). Ces descripteurs permettent de calculer des propriétés de la molécule qui concernent tant ses caractéristiques élémentaires (types d'atomes, groupements fonctionnels,...) que des caractéristiques topologiques et géométriques plus complexes issues notamment de la structure 3D des molécules. Les descripteurs Dragon permettent ainsi de recouvrir une grande diversité de propriétés des molécules étudiées. Nous avons utilisé E-Dragon (http://www.vcclab.org/lab/edragon/), la version en ligne de ce logiciel qui permet de calculer 1666 descripteurs différents. Nous avons alors pré-sélectionné les variables afin d'éliminer les variables constantes ou redondantes (coefficient de corrélation linéaire supérieur à 0.9). Nous avons ainsi obtenu 357 descripteurs qui ont été utilisés pour construire les modèles.

2.2 Echantillon de validation

Nous avons extrait 26 nouveaux inhibiteurs de PPI. Pour vérifier que ce jeu n'est pas redondant avec les inhibiteurs utilisés pour le jeu d'apprentissage, nous avons calculé la distance de Tanimoto (Robert and Carbó-Dorca, 1998). Seulement 2 des 26 inhibiteurs de validation ont un indice de Tanimoto supérieur 0.8 (mais restant inférieur à 0.9) avec une des molécules du jeu d'apprentissage. On a donc bien un jeu constitué de molécules nouvelles par rapport à l'apprentissage. Les molécules *négatives* ont été obtenues à partir d'une autre base de données le *ChemBridge diversity set*. Les mêmes filtrages ADMET et diversité ont été appliqués amenant à conserver 2000 molécules.

3 Construction du modèle de classification

3.1 Choix de la méthode de discrimination

Comme souvent lorsque l'on utilise une méthode statistique dans un contexte de recherche appliquée, deux objectifs doivent la plupart du temps être pris en compte :

- la précision du modèle : il doit permettre de prédire avec exactitude la ou les caractéristiques souhaitées,
- l'interprétabilité du modèle : les utilisateurs doivent pouvoir utiliser ce modèle pour leur fournir une meilleur compréhension du phénomène sous-jacent.

La deuxième partie est optionnelle si la prédiction prime sur la compréhension. Toutefois, un modèle permettant d'atteindre ces deux objectifs est encore plus intéressant.
Dans ce but, nous avons commencé par utiliser des arbres de classification et également une méthode plus complexe susceptible de fournir des modèles plus précis bien que très peu interprétables, la méthode des Séparateurs à Vastes Marges (SVM).

3.2 Optimisation des arbres de classification

Afin d'adapter au mieux la méthode des arbres de classification à notre problématique, nous avons utilisé, comme critère de division, l'enrichissement permis par la séparation d'un nœud. Ce critère se définit comme le taux de vrais positifs parmi les observations déclarées comme positives à l'issue d'une division par un nœud, rapporté au taux de vrais positifs présents dans l'échantillon avant la division. Il s'agit donc d'une valeur réelle positive qui est d'autant plus grande que la division choisie a permis un enrichissement important de l'échantillon déclaré comme *positif* à l'issue de la division. Ce critère est ainsi adapté à l'objectif de proposer un ensemble de composés plus riche en inhibiteurs de PPI que la banque de départ.
Afin d'éviter le sur-ajustement, 20 arbres ont été construits par validation croisée (chaque arbre est construit sur 19/20 des observations initiales). L'arbre final choisi est donc un consensus des différents arbres obtenus en ne retenant que les divisions apparaissant le plus souvent à travers les 20 arbres obtenus et tenant compte des résultats de sensibilité, spécificité et enrichissement.

3.3 Optimisation des SVM

Trois noyaux ont été testés : gaussien, polynomial, tangente hyperbolique. Les paramètres de chaque noyau ainsi que la constante permettant de contrôler le compromis entre taux d'erreurs et largeur de la marge ont été optimisés en validation croisée. Les combinaisons de paramètres finales ont été choisies de sorte à réaliser un compromis entre sensibilité, spécificité et enrichissement.

4 Résultats

En ce qui concerne les arbres de classification, deux modèles ont finalement été retenus, l'un privilégie la sensibilité (D.T. 1), l'autre la spécificité (D.T. 2) comme l'indique la Table 1. Nous avons retenu deux arbres afin de pouvoir s'adapter à différents objectifs d'application : si l'on privilégie la réduction de la taille de la base de données on choisira plutôt l'arbre D.T. 2 alors que si on souhaite limiter le risque de perte d'inhibiteurs de PPI, on optera pour l'arbre D.T. 1. Les résultats pour les différents noyaux en SVM y sont également présentés.

On constate globalement que les SVM ont de très bons résultats sur l'échantillon d'apprentissage mais que la perte de performance est très importante en validation. Ainsi, malgré l'optimisation des paramètres en validation croisée, pour ces données, les méthodes SVM ont une tendance importante au sur-ajustement. Ce défaut est nettement moins important pour les arbres de décision où les performances sont assez conservées entre jeux de données.

Table 1: Résultats obtenus sur le jeu d'apprentissage (Ap), en validation croisée sur ce jeu (CV) et sur le jeu de validation (Val) pour les différentes méthodes utilisées (D.T. : arbre de classification et SVM). Les indicateurs présentés sont la sensibilité (Se), la spécificité (Sp) et l'enrichissement (E).

Method	D.T. 1		D.T. 2		SVM gaus			SVM tanh			SVM poly		
Dataset	Ap	Val	Ap	Val	Ap	CV	Val	Ap	CV	Val	Ap	CV	Val
Se (%)	85	81	76	70	89	39	33	92	42	33	89	33	27
Sp (%)	70	66	77	80	100	97	85	100	93	81	100	98	84
E	2.38	2.53	2.61	3.39	9.44	5.71	2.23	9.29	3.83	1.77	9.44	5.77	1.67

L'intérêt des arbres de classification est double puisqu'outre leurs performances plus importantes en validation, ils fournissent un modèle interprétable qui permet une meilleure compréhension du phénomène. Ainsi la Figure 1 montre les deux arbres finaux. On y voit l'intervention de deux descripteurs, RDF070m et Ui. RDF070m est une fonction de distribution radiale pondérée par la masse atomique des atomes en utilisant une sphère de rayon 7 Å. Elle peut être interprétée comme la probabilité de trouver un atome dans une sphère et reflète la forme de la molécule : plus la molécule est allongée plus la valeur de RDF070m est faible et inversement, plus la molécule est ramifiée (en forme de T ou d'étoile par exemple), plus cette valeur est forte. Il semble donc que certains PPI se caractérisent par une forme plutôt ramifiée. D'autre part, Ui (Unsaturation Index) est une fonction logarithmique du nombre de liaisons multiples dans la molécule. Les deux arbres diffèrent justement par le seuil affecté à ce descripteur : D.T. 1 correspond à un seuil de 15 liaisons multiples et D.T. 2 à un seuil de 17 liaisons multiples. Les résultats ont aussi été validés sur d'autres banques de données (résultats non présentés ici).

5 Discussion et Conclusion

Le modèle construit lors de cette étude est très important pour la recherche dans le domaine des inhibiteurs de PPI. En effet, il a eu le mérite de montrer qu'il était possible de construire un modèle global qui permet de discriminer les inhibiteurs de PPI d'autres types de molécules thérapeutiques quelle que soit la cible de ces inhibiteurs. L'espace chimique des inhibiteurs de PPI peut donc bien se distinguer de l'ensemble des molécules thérapeutiques *classiques*, et ce, à l'aide de seulement deux descripteurs. De plus, le modèle nous a apporté des connaissances intéressantes sur le profil-type des inhibiteurs de PPI puisqu'ils sont plutôt ramifiés et contiennent au moins 15 ou 17 liaisons multiples. Ce travail a donné lieu à la création d'un programme, disponible gratuitement sur demande à l'adresse www.CDithem.com, qui permet aux utilisateurs de filtrer une banque de données pour extraire les molécules susceptibles d'être des inhibiteurs de PPI.

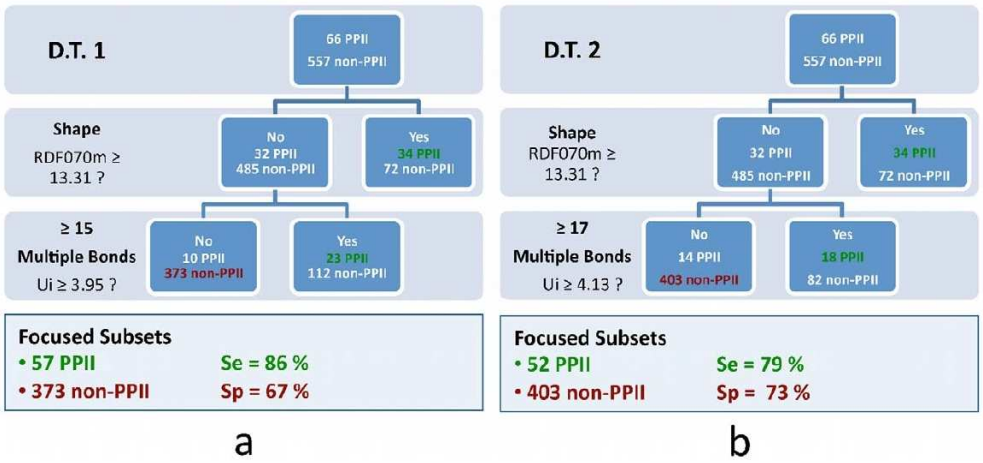


Figure 1: Structure des arbres finaux D.T. 1 et D.T. 2. et performances sur le jeu d'apprentissage.

References

Arkin, M. and Whitty, A. (2009). The road less traveled: modulating signal transduction enzymes by inhibiting their protein-protein interactions. *Current opinion in chemical biology*, 13(3):284–290.

Reynès, C., Host, H., Camproux, A., Laconde, G., Leroux, F., Mazars, A., Deprez, B., Fahraeus, R., Villoutreix, B., and Sperandio, O. (2010). Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors using Machine-Learning Methods. *PLoS Computational Biology*, 6(3):6959–6964.

Robert, D. and Carbó-Dorca, R. (1998). A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comput. Sci*, 38(3):469–475.

Sperandio, O., Reynès, C., Camproux, A., and Villoutreix, B. (2010). Rationalizing the chemical space of protein-protein interaction inhibitors. *Drug discovery today*, 15(5-6):220–229.

Stumpf, M., Thorne, T., De Silva, E., Stewart, R., An, H., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959.

Todeschini, R., Consonni, V., and MyiLibrary (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH Weinheim, Germany.

Classification des données quantitatives de grande dimension dans l'environnement logiciel MIXMOD

Christophe Biernacki[a], Gilles Celeux[b], Gérard Govaert[c], Florent Langrognet[d]

[a] *Laboratoire Paul Painlevé (Univ. de Lille 1. - CNRS) – 59 655 Villeneuve d'Ascq*
[b] *Projet Select (INRIA Saclay Île de France) – Université Paris-Sud – 91405 ORSAY*
[c] *Laboratoire Heudiasyc (Univ. de Technologie de Compiègne - CNRS) – 60205 COMPIÈGNE*
[d] *Laboratoire de Mathématiques de Besançon (Univ. de Franche-Comté - CNRS) – 25030 BESANÇON*

Résumé L'ensemble logiciel MIXMOD (MIX*ture* MOD*elling*) permet de traiter des problématiques de classification supervisée et non supervisée de données quantitatives ou qualitatives dans un contexte de modèle de mélange (Biernacki *et al.* 2006). 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. MIXMOD se compose d'une bibliothèque de calcul robuste et performante et d'outils complémentaires : des fonctions pour Matlab et une interface graphique (mixmodGUI).

Mots clés Modèles gaussiens et multinomiaux, algorithmes EM, sélection de modèles, grande dimension.

Abstract The software MIXMOD (MIX*ture* MOD*elling*) allows to deal with clustering and classification for continuous and discrete data sets by using mixture model (Biernacki *et al.* 2006). Several mixture parameter estimation algorithms are available (EM, CEM, SEM) and combining them is possible to get efficient initialisation strategies. Penalised likelihood criteria to select a parsimonious model and a sensible number of mixture parameters are available. Moreover, Gaussian mixture for continuous data and multivariate multinomial distributions for discrete data are available, and recently specific Gaussian mixture for high dimensional data have been included. The MIXMOD sofware is a reliable and efficient data analysis library interfaced with Matlab and a graphical user interface is now available (mixmodGUI).

Keywords Multivariate Gaussian and multinomial mixtures, EM-like algorithms, model selection, high dimensional data.

1 Présentation générale

Les modèles de mélange finis sont un puissant outil pour la classification supervisée et non supervisée des données (voir par exemple McLachlan et Peel 2000 pour une vue synthétique).

MIXMOD propose des modèles de mélange gaussiens multivariés pour les données quantitatives et des mélanges de lois multinomiales multivariées (modèle des classes latentes) pour les données qualitatives.

Depuis deux ans, MIXMOD propose également des modèles spécifiques pour traiter des données de grande dimension dans le cadre supervisé. En 2011, ces modèles permettront également de traiter les problématiques de classification non supervisée.

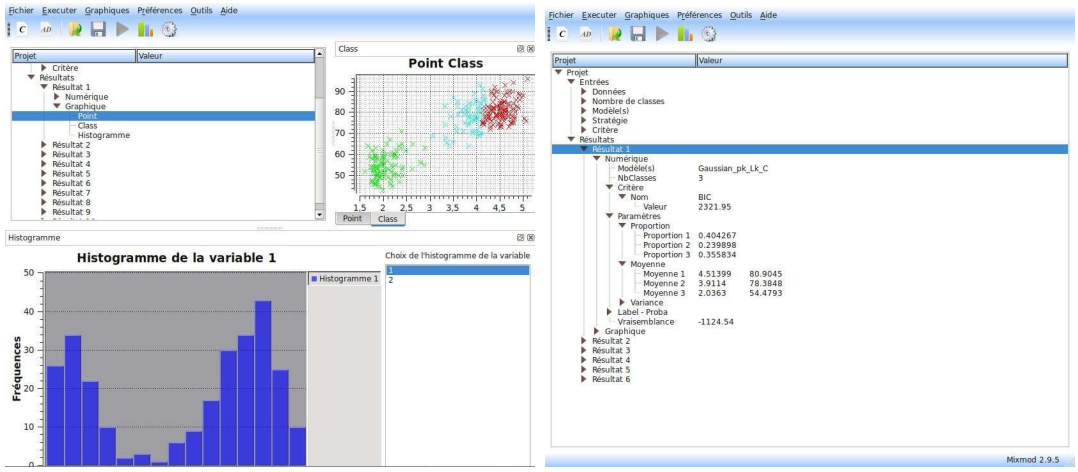
Cet ensemble logiciel est destiné aux scientifiques et aux ingénieurs qui ont besoin de résoudre des problèmes de classification ou de discrimination. Dans un contexte universitaire, il s'agit d'un outil d'enseignement pour des cours d'introduction ou avancés dans ces domaines. Dans l'industrie, MIXMOD est un instrument de choix pour la recherche, l'analyse et en général la reconnaissance statistique des formes.

2 L'ensemble logiciel MIXMOD

Le projet MIXMOD est le fruit d'un partenariat débuté en 2001. Christophe Biernacki (professeur au laboratoire Paul Painlevé de Lille), Gilles Celeux (directeur de recherche à l'INRIA Saclay Île de France), Gérard Govaert (professeur au laboratoire Heudiasyc, Compiègne) et Florent Langrognet (ingénieur de recherche au laboratoire de mathématiques de Besançon) forment le comité de pilotage du projet intégrant des compétences en statistiques et en informatique. L'objectif est de développer et diffuser un ensemble logiciel de classification de données qui soit à la fois puissant, rapide et robuste : une bibliothèque de calcul écrite en C++ (70 classes et 40 000 lignes de code) intégrant toutes les fonctionnalités et des outils pour utiliser MIXMOD. À ce jour, il s'agit :

- de fonctions pour le logiciel MATLAB ;
- d'une interface graphique (mixmodGUI) basée sur la bibliothèque graphique QT (voir Figure 1).

Ainsi, les utilisateurs d'horizons différents disposent d'outils complémentaires pour accéder à la richesse de la bibliothèque de calcul MIXMOD. De plus, il est prévu de développer des fonctions pour le logiciel R dans les prochains mois pour répondre à une demande spécifique pour cet environnement.



(a) (b)

Figure 1: visualisation des résultats (a) et des résultats numériques (b) avec mixmodGUI

Le site web de MIXMOD (www.mixmod.org) est le portail du projet. Toutes les informations concernant l'ensemble logiciel sont disponibles (téléchargement, documentations (statistique, userguide), news, ...). On y trouve également un forum de discussion et une rubrique regroupant les articles et communications traitant de MIXMOD. La fréquentation de ce site web est d'environ 700 visites par mois depuis des années et on compte entre 200 et 250 téléchargements par mois. La diffusion de MIXMOD est favorisée par le choix des licences qui sont associées aux différents produits :

- licence GNU GPL pour la bibliothèque de calcul et les fonctions pour Matlab (et bientôt R) permettant d'adapter le code à d'éventuels besoins spécifiques ;
- licence propriétaire et gratuite pour l'interface graphique mixmodGUI.

L'équipe MIXMOD organise tous les deux ans une rencontre avec les utilisateurs (actuels et potentiels). Après Paris en 2006, Lille en 2008, la 3^e rencontre MIXMOD a eu lieu à Lyon en décembre 2010. Ces rencontres sont l'occasion, d'une part, de présenter MIXMOD, ses nouveautés et les perspectives et, d'autre part, de donner la parole aux utilisateurs pour des présentations d'utilisations concrètes de MIXMOD.

3 La classification des données quantitatives de grande dimension avec MIXMOD

Nous avons implémenté une famille de huit modèles gaussiens spécifiques de la haute dimension. Ces modèles introduits par Bouveyron *et al.* (2007) sont l'une des variantes

de mélange d'analyse factorielle (voir aussi McNicholas and Murphy 2008) et utilisent une reparamétrisation des matrices variances des composants Σ_k, $\forall k = 1, ..., K$ fondée sur leur décomposition spectrale

$$\Sigma_k = Q_k \Delta_k Q_k^t,$$

Q_k étant la matrice orthogonale des vecteurs propres de Σ_k et Δ_k la matrice diagonale de ses vaeurs propres ayant la forme suivante:

$$\Delta_k = \left(\begin{array}{cc} \boxed{\begin{array}{ccc} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{k\delta_k} \end{array}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{array}{ccc} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{array}} \end{array} \right) \begin{array}{l} \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \delta_k \\ \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} (d - \delta_k) \end{array}$$

où $a_{kj} \geq b_k$, pour $j = 1, ..., \delta_k$ et $\delta_k < d$. Le sous-espace de la classe k engendré par les premières valeurs propres a_{kj}, contenant la moyenne μ_k est noté $\mathbb{E}_k$. Dans l'espace orthogonal à $\mathbb{E}_k$, la variance de la classe est caractérisée par un seul paramètre b_k. Les projecteurs sur $\mathbb{E}_k$ et $\mathbb{E}_k^\perp$ sont notés P_k et $P_k^\perp$. Partant du modèle général $[a_{kj} b_k D_k \delta_k]$, huit modèles particulièrement utiles ont été considérés dans MIXMOD:

- Deux modèles ont des dimensions libres δ_k:
  - modèle $[a_{kj} b_k D_k \delta_k]$
  - modèle $[a_k b_k D_k \delta_k]$
- Six modèles ont une dimension fixe δ:
  - modèle $[a_{kj} b_k D_k \delta]$
  - modèle $[a_j b_k D_k \delta]$
  - modèle $[a_{kj} b D_k \delta]$
  - modèle $[a_j b D_k \delta]$
  - modèle $[a_k b_k D_k \delta]$
  - modèle $[a_k b D_k \delta]$

Leur estimation dans un cadre supervisé ou non par le maximum de vraisemblance sera décrit dans cet exposé. On montrera que cette estimation ne souffre pas des défauts numériques inhérents à l'estimation en haute dimension et on illustrera le fait que, dans le cadre supervisé en haute dimension, ils fournissent des modèles plus parcimonieux que l'analyse discriminante linéaire bien que non linéaires et plus réalistes. Enfin, on discutera d'heuristiques pour un choix efficace de la dimension intrinsèque δ_k.

4 Perspectives

L'une des vocations de MIXMOD est de traiter des donneées dans des contextes difficiles, notamment des données de grande taille. Aussi, nous comptons proposer dans MIXMOD les algorithmes développés par Govaert et Nadif (2008) pour l'analyse du modèle des blocs latents qui permet une classification en blocs parcimonieux de grands ensembles de données, comportant de très nombreuses lignes et de très nombreuses colonnes.

Sur le plan informatique, les prochains travaux porteront sur :

- le développement d'un package MIXMOD pour R ;
- l'enrichissement de l'interface graphique mixmodGUI ;
- l'adapatation du package MIXMOD pour Matlab ;
- la poursuite de la recherche des meilleures performances pour la bibliothèque de calcul MIXMOD.

Bibliographie

[1] G.J. McLachlan et D. Peel. (2000), *Finite Mixture Models*. New York: Wiley.
[2] C. Biernacki, G. Celeux, G. Govaert et F. Langrognet (2006), Model-based cluster analysis and discriminant analysis with the MIXMOD software, *Computational Statistics and Data Analysis*, 51, 587-600.
[3] C. Bouveyron, S. Girard et C. Schmid (2007) High-Dimensional Data Clustering, *Computational Statistics and Data Analysis*, 52, pp. 502-519.
[4] G. Govaert et M. Nadif (2008) Block clustering with Bernoulli mixture models: Comparison of different approaches, *Computational Statistics and Data Analysis*, 52, 3233-3245.
[5] P.D. McNicholas et T.B. Murphy (2008) Parsimonious Gaussian mixture models, *Statistics and Computing*, 18, 285-296.

ELABORATION D'UN AGE BIOLOGIQUE A PARTIR DE DONNEES ACCESSIBLES EN ROUTINE DE MEDECINE GENERALISTE

Essai de fondement théorique

M. Sarazin[1], X. Bay[2]

RESUME

Le vieillissement caractérise une évolution inéluctable du corps dont la quantification est établie par l'âge dépendant du temps dit « chronologique ». Cependant, ce critère âge ne quantifie qu'imparfaitement l'usure réelle du corps soumise à de nombreux facteurs modificateurs dépendant des individus. Aussi, a-t-il été substitué depuis longtemps [1-9] par un critère composite, appelé « âge biologique », sensé davantage refléter le vieillissement individuel. Afin d'essayer d'en faire un outil quantificateur accessible à la pratique de médecine générale, une nouvelle méthodologie est proposée.

METHODOLOGIE

Le critère « âge biologique » a été défini à partir de la grandeur âge « chronologique » et de variables cliniques et biologiques pondératrices caractérisant l'état de santé du corps humain et mesurables au cours d'un examen médical standard. Un échantillon de population témoin supposé « vieillir normalement », selon les critères de normalité des variables utilisées, a servi pour le calcul du facteur pondérateur de chacune des variables. Le sexe a été au préalable fixé. La dépendance statistique des variables utilisées a été modélisée par une copule gaussienne [10-12] (prise en compte seulement de corrélations linéaires deux à deux). Le modèle est donc le suivant pour les variables après normalisation :

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon = X^T\beta + \varepsilon$$

où Y est l'âge chronologique d'un individu, et X_1, …, X_p les p variables biologiques mesurées et où le résidu ε (nécessairement centré) est indépendant du vecteur $X = (X_1, \ldots, X_p)^T$. Les coefficients β sont solution du système linéaire $\Gamma_p\beta = b$ où $\beta = (\beta_1, \ldots, \beta_p)^T$, Γ_p est la matrice de corrélation des variables X_k et $b = (\mathrm{cov}(Y, X_1), \ldots, \mathrm{cov}(Y, X_p))^T$.

L'âge biologique est défini de la manière suivante

$$Z_p = Y + \sum_{k=1}^{p} \beta_k(X_k - \alpha_k Y)$$

[1] *1. INSERM, U707, F-75012 Paris, France 2. Centre Hospitalier, Rue de Bénaud, 42700 Firminy*

[2] *Ecole Nationale Supérieure des Mines de Saint-Etienne, Institut Fayol, 158 cours Fauriel, F-42023 SAINT-ÉTIENNE*

où les coefficients α_k sont donnés par les régressions linéaires : $X_k = \alpha_k Y + \varepsilon_k$, $1 \leq k \leq p$.

La variable âge biologique Z_p est estimée à partir des p variables biologiques X_1, …, X_p , et par extension , la variable $IV_p = Z_p - Y$ est appelé ***indicateur de vieillissement*** (estimé à partir de X_1, …, X_p).

Les coefficients α_k vérifient la relation : $\Gamma_p \beta = \mathrm{Var}(Y) \times \alpha$ où $\alpha = (\alpha_1, \ldots, \alpha_p)^T$. Il est déduit que $IV_p = Z_p - Y = 0$ lorsque $Y = \beta_1 X_1 + \ldots + \beta_p X_p$ (variance résiduelle nulle ou coefficient R^2 maximal). La validité théorique du modèle est donc établie lorsque l'âge chronologique est entièrement expliqué par les variables biologiques.

Sur un plan statistique, les coefficients du modèle sont estimés à partir de la population vieillissant normalement, ce qui nécessite au préalable l'estimation des lois marginales de toutes les variables. Mais, dans l'utilisation pratique du critère âge biologique ou de l'indicateur de vieillissement pour un individu quelconque, il est indispensable de bien connaître ces lois marginales au-delà des quantiles usuels (intervalle à 95%), c'est-à-dire de bien connaître la distribution de leurs valeurs extrêmes renforçant la discrimination du modèle. La méthode des excès est utilisée pour cela [13].

PERSPECTIVES

Cette méthode ouvre de nouvelles perspectives en termes de prévention et prise en charge du risque de vieillissement. Cependant, la pertinence du modèle doit être validée par des études de morbidité et un retour d'expérience des médecins généralistes.

ABSTRACT

Aging characterized unavoidable changes in the body. Its measurement is commonly determined by the age dependent on time and called “chronological age”. However, the criterion « chronological age » reflects imperfectly the actual aging of the body depending on many individual factors. Also, this criterion has since long been replaced by another composite criterion called « biological age » [1-9] supposed to better reflect the aging process. In order to build a score of aging adapted to general practice, new ways of thinking are proposed.

METHODOLOGY

The criterion "biological age" was defined from the quantity “chronological” age and, clinical and biological variables characterizing the health of the human body and measurable in a standard medical examination. A control population sample supposed of "normal aging ", according to the criteria of normality of variables, was used for calculating the weighting role of each variable. Gender has been previously fixed. The statistical dependence of variables has been modeled by a Gaussian copula [10-12] (taking into account only linear correlations of pairs). The model is as follows for the variables after normalization:

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon = X^T\beta + \varepsilon$$

where Y is the chronological age of an individual, and X1, ..., Xp p biological variables measured and where the residue □ (necessarily centered) is independent of the vector X = (X1, ..., Xp) T. The vector of coefficient □ is solution of the linear system $\Gamma\ \beta$ p = b where β = (β 1, ..., β p) T, Γ p is the correlation matrix of variables X k and b = (cov (Y, X1), ..., cov (Y, Xp)) T.

“Biological age” is defined as follows:

$$Z_p = Y + \sum_{k=1}^{p} \beta_k (X_k - \alpha_k Y)$$

Where α_k coefficients are given by the linear regressions : $X_k = \alpha_k Y + \varepsilon_k$, $1 \leq k \leq p$.

“Biological age”, Zp, is estimated from the p biological variables X1, ..., Xp, and by extension, variable = Zp IVp – Y is called ***“aging indicator”*** (estimated from X1, ..., Xp).

α k coefficients satisfy the relation: p $\Gamma\ \beta$ = Var (Y) × α where α = (α 1, ..., α p) T. By deduction, IVp = Zp – Y = 0 when $Y = \beta_1 X_1 + \ldots + \beta_p X_p$ (zero residual variance or coefficient R2 maximum). Hence, the consistency of the model is established when chronological age is entirely explained by biological and clinical variables.

On a statistical way, coefficients of the model are estimated from the sample population aging normally. It requires first to estimate the marginal distributions of all variables. But in the practical use of the criterion “biological age” or “aging indicator” for any individual, it is essential to be familiar with these marginal distributions beyond the usual quantiles (range 95%). Indeed, it is necessary to have a real knowledge of the distribution of extreme values to increase the discriminating power of the model. The method of excess is used [13].

PERSPECTIVES

New perspectives are offered in terms of prevention and management of the risk of aging. However, the relevance of the model must be validated by studies of morbidity and feedback from general practitioners.

REFERENCES

1. *Uttley M, Crawford MH. .Efficacy of a composite biological age score to predict ten year survival among kansas and nebraska Mennonites. Hum Biol 1994 Feb, 66 (1):121-44*

2. *Ruiz-Torres A, Agudo A, Vicent D, Beier W. Measuring human aging using a two compartimental mathematical model and the viality concept. Arch Gerontol Geriatr 1990 Jan Feb; 10 (1) : 69-76*

3. *Jazwinski SM. Biological aging research today: potential, peeves and problems. Exp Gerontol, 2002 Oct-nov; 37(10-11):1141-6*

4. *Song X, Skoog I, Broe GA, Cox JL, Grunfeld E, Rockwood K. Relative fitness and frailty of elderly men and women in developed countries and their relationship with mortality. J Am Geriatr Soc. 2005 Dec;53(12):2184-9*

5. *Linda P. Fried, Catherine M. Tangent et al. Frailty in older adults: evidence for a phenotype. J of gerontology, 2001, vol 56A, N°3, M146-M156*

6. *Varadhan R., Seplaki CL. et al. Stimulus-response paradigm for characterizing the loss of resilience in homeostatic regulation associated with frailty. Mech Ageing Dev. 2008 Nov;129(11):666-70. Epub 2008 Sep 30*

7. *Yates FE. Modeling "frailty": can a simple feedback model suffice?. Mech Ageing Dev. 2008 Nov;129(11):671-2. Epub 2008 Oct 5.*

8. *Geguen R. .Proposition of an aging indicator from general health examination in France. Clin Chem Lab Med 2002; 40(3): 235-239*

9. *Bandeen roche K., Xue QL. Phenotype of frailty : characterization in the women's health and aging studies. Journal of gerontology, 2006, vol 61A, n°3, 262- 266*

10. *Scherrer B. Biostatistiques. Volume 1. Editions G Morin . 2007: paragraphes 16.2.3, 8.2.6.1, 8.2.6.1, 8.2.6.2*

11. *Legendre P. , Legendre L. Numerical ecology; developments in environmental modelling. Elsevier Science Amsterdam. 1998: 178.*

12. *Nelsen R. B. An Introduction to Copulas. Springer, New York.1999*

13. *Kotz S, Nadarajah S. Extreme value distributions, theory and applications. Imperial College Press. 2000*

L'APPROCHE PLS POUR LA RECHERCHE DE MARQUEURS DANS LE CADRE D'UNE ETUDE CLINIQUE OBSERVATIONNELLE EN NUTRITION

Marie Keravec, Pascale Rondeau, Sébastien Vergne, Sébastien Marque

Danone Research, Centre Daniel Carasso, RD128, Avenue de la Vauve, F-91767 Palaiseau Cedex

Mots clés: recherche de marqueurs, étude observationnelle, PLS

Résumé

Lors d'une étude clinique observationnelle en nutrition, de nombreuses données de natures différentes sont collectées : recueils alimentaires, données provenant de différentes matrices biologiques. L'un des objectifs de ce type d'étude est d'identifier des biomarqueurs liés à la consommation alimentaire.

Afin de répondre à cet objectif, l'approche PLS est évaluée sur une matrice de données comprenant quatre-vingt-deux sujets, quatre-vingt-deux paramètres biologiques et la consommation journalière de fluides.
Les sujets sont répartis en trois catégories de consommation : trente-neuf petits buveurs (sujets déclarant des consommations de fluides inférieures à 1,2L/jour), onze buveurs moyens (consommations entre 1,2 et 2L/jour) et trente-deux gros buveurs (consommations supérieures à 2L/jour).

Un modèle PLS1 conduit à l'identification d'une dizaine de marqueurs urinaires cependant, l'erreur d'estimation du modèle est élevée (±597 mL) par rapport à la quantité moyenne de fluide déclarée par jour (environ 1600 mL). Cette imprécision d'estimation peut être due à la grande variabilité de la consommation de fluides (Y) par rapport à une faible variabilité des marqueurs biologiques sélectionnés (Xj).

Au vu de ces résultats, une PLS DA a été menée afin d'évaluer la robustesse des marqueurs identifiés. Afin de s'affranchir de l'hypothèse de linéarité du modèle PLS, la méthode de classification supervisée random forest a également été appliquée.
Les résultats issus de ces méthodes seront présentés et discutés.

.

PLS APPROACH APPLIED TO MARKERS RESEARCH IN A NUTRITION OBSERVATIONAL STUDY

Marie Keravec, Pascale Rondeau, Sébastien Vergne, Sébastien Marque

Danone Research, Centre Daniel Carasso, RD128, Avenue de la Vauve, 91767 Palaiseau Cedex

Keywords**:** marker research, observational study, PLS model

Abstract

During an observational clinical study in nutrition, many data from various sources are collected: nutritional data and data from various biological matrices. One of the objectives of this type of study is to identify biomarkers linked to food consumption.

In order to answer this objective, the PLS approach is tested on a matrix of data including eighty-two subjects, eighty-two biological parameters and the daily consumption of fluids.
Subjects are distributed in three categories of consumption: thirty-nine low drinkers (declaring consumptions of fluids lower than 1,2L / day), eleven medium drinkers (consumptions between 1,2 and 2L per day) and thirty-two high drinkers (consumptions superior to the 2L / day).

A PLS1 model leads to the identification of about ten urinary markers, however, the error of estimation of the model is high (±597 mL) compared to the mean quantity of fluid consumed per day (around 1600 mL). This inaccuracy of estimation can be due to the high variability of the consumption of fluids (Y) combined to a low variability of the selected biological markers (Xj).

Considering these results, a PLS DA was performed to estimate the robustness of the identified markers. To resolve the hypothesis of linearity of the PLS model, the method of supervised classification random forest was also applied.
The results from each method will be presented and discussed.

On stationarity and existence of moments of the spatial RCA models

Karima Kimouche

Département de Mathematiques,
Université Mentouri Constantine, Algeria.
e-mail: karima_dino@yahoo.fr

1 Introduction

Pendant les deux dernières décennies, il y a eu un intérêt croissant pour les modèles non-linéaires, cet intérêt est motivé de faire savoir que les modèles non-linéaires sont souvent produire de meilleures prévisions que les modèles linéaires. L'un de ces modèles qui ils ont reçu l'attention considérable sont les modèles RCA. A l'origine, ces modèles ont été inventés dans le contexte des perturbations aléatoires des systèmes dynamiques, mais maintenant ils ont utilisé dans divers applications, par exemple, dans la finance et la biologie [8].

Ainsi, par l'extension de certains modèles non linéaire d'un dimension à multiple dimension, les modèles autorégressif à coefficient aléatoire spatiaux sont la généralisation des modèles autorégressifs spatiaux qui ils ont été étudier par plusieurs auteurs, par exemple, Tjosteim [9], Choi [3], Guyon [5].

A deux dimensions, nous intéresserons notre attention sur les modèles $(SRCA)$ d'ordre (p_1, p_2) comme

$$X(i,j) = \sum_{k=0}^{p_1}\sum_{l=0}^{p_2}[\beta_{kl} + \beta_{kl}(i,j)]X(i-k, j-l) + \varepsilon(i,j), \quad \beta_{00} = 0, \beta_{00}(i,j) = 0 \ . \qquad (1)$$

où $(X(i,j))_{(i,j)\epsilon Z^2}$ est un processus spatial de valeurs à R sur une grille régulièrement rectangulaire définie sur un espace de probabilité $(\Omega, \Im, P)$. On définis sur Z^2, $Z := \{0, \pm 1, \pm 2, ...\}$ l'ordre lexicographique i.e., pour $\mathbf{s} = (s_1, s_2)$ et $\mathbf{t} = (t_1, t_2) \in Z^2$, on écrit $\mathbf{s} \preceq \mathbf{t}$ si l'un $(s_1 < t_1)$ ou $(s_1 = t_1$ et $s_2 < t_2)$.

nous faisons les hypothèses suivantes

i) $\{\varepsilon(i,j); (i,j)\epsilon Z^2\}$ est une suite de variables aléatoires indépendantes et identiquement distribué $(i.i.d.)$ de moyenne nulle et de variance σ^2.

ii) Les $\beta_{kl}, k = 0, ..., p_1, l = 0, ..., p_2$, sont des constantes réelles.

iii) $\{\beta_{kl}(i,j); (i,j)\epsilon Z^2\}$ est une suite de variables aléatoires indépendantes et identiquement distribué $(i.i.d.)$ de moyenne nulle et de variance w^2, et $(\beta_{kl}(i,j))_{0<k<p_1, 0<l<p_2}$ est indépendante de $(\varepsilon(i,j))$ pour tout $(i,j)\epsilon Z^2$.

Le papier est organisé comme suit. Dans la section 2, nous donnons des conditions ncessaires et suffisantes pour l'existence d'une solution stationnaire et ergodique de l'quation (1). La condition assurant l'existence de moments d'ordre supérieur est donné dans la section 3. Enfin, la covariance, la densité spectrale pour $SRCA(1)$ sont obtenus.

2 Stationarité

Nous utilisons la notation suivante. Soit $x = [X(i,j)]$ une matrice $(p_1+1)(p_2+1)$ d'observation sur les series spatiales dans le plan .

Nous posons $\underline{X}(i,j) = vec(x)$ qui désigne le $P \times 1$ vectorisation de la matrice x avec $P = (p_1+1)(p_2+1)$. Alors

$$\underline{X}(i,j) = \begin{bmatrix} X_i & X_{i-1} & \cdots & X_{i-p_1} \end{bmatrix}' \tag{2}$$

où $X_{i-k} = (X(i-k,j), X(i-k,j-1), ..., X(i-k,j-p_2))', k = 0, ..., p_1$. Ces impliquent que nous pouvons écrire l'quation (1) sous la forme matricielle

$$\underline{X}(i,j) = [W + \Pi(i,j)]\,\underline{X}(i,j) + \underline{\varepsilon}(i,j) \tag{3}$$

où W$= \begin{bmatrix} W_0 & W_1 & \cdots & W_{p_1} \\ 0 & I_{p_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & I_{p_2} \end{bmatrix}, \Pi(i,j) = \begin{bmatrix} \Pi_0(i,j) & \Pi_1(i,j) & \cdots & \Pi_{p_1}(i,j) \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix},$

$\underline{\varepsilon}(i,j) = \begin{bmatrix} \varepsilon_i & 0 & \cdots & 0 \end{bmatrix}'$. Notons que $\underline{X}(i,j)$, $\underline{\varepsilon}(i,j)$ sont des vecteurs de $P \times 1$ dimension, tandis que W et $\Pi(i,j)$ sont des matrices carrées de $P \times P$ dimension avec $E[\varepsilon_i \varepsilon_i'] = \zeta$, et $E[\Pi_k(i,j) \otimes \Pi_k(i,j)] = \Gamma_k$.

Notre resultat principal est contenus que le modèle autorgressif à coefficient aléatoire spatial (3) satisfait les hypothses (i) - (iii) a une solution unique stationnaire et ergodique . Cette solution est donnée par

$$X_i = \sum_{n=0}^{\infty} T_i^n \varepsilon_{i-n}, \tag{4}$$

où la matrice de transition T_i^n est définis par

$$T_i^n = \sum_{k=0}^{p_1} (W_k + \Pi_k) T_i^{n-k}, T_i^0 = I_{p_2}. \tag{5}$$

3 Existence de moments

Dans cette section, nous donnons la condition d'existence de moments d'ordre supérieur pour le processus $\underline{X}(i,j)$ définie par (3), qui est convaincu les hypothèses ci-dessus. Nous

concluons par un exemple d'un processus $SRCA_2(1,1)$ pour illustrer

$$X(i,j) = [\beta + \beta(i,j)]X(i-1,j-1) + \varepsilon(i,j). \tag{6}$$

Nous posons que le processus $(X(i,j))_{(i,j)\in Z^2}$ est stationaire du second-order, et nous obtenons sa fonction de covariance $C(h_1,h_2)$ et sa densité spectrale f(λ).

4 Introduction

During the past two decades, there has been a growing interest in nonlinear models, this interest is motivated namely by the fact that nonlinear models are often produce better forecasts than a linear models. One of the models which received considerable attention is the RCA models. Originally, these models have been invented in the context of random perturbations of dynamical systems, but are now used in a variety of application for example, in finance and biology [8].

Hence, by extending some one-dimensional nonlinear models to multiple dimension one, spatial random coefficient autoregressive models are generalization of spatial autoregressive models which have been heavily studied by a several authors for instance, Tjosteim [9], choi [3], Guyon [5].

In two dimensions, we shall focus our attention on $(SRCA)$ models of order (p_1,p_2) as

$$X(i,j) = \sum_{k=0}^{p_1}\sum_{l=0}^{p_2}[\beta_{kl} + \beta_{kl}(i,j)]X(i-k,j-l) + \varepsilon(i,j), \quad \beta_{00} = 0, \beta_{00}(i,j) = 0\ . \tag{7}$$

where $(X(i,j))_{(i,j)\epsilon Z^2}$ is a R-valued spatial process on a regular rectangular grid defined on some probability space $(\Omega, \Im, P)$. we define on Z^2, $Z := \{0, \pm 1, \pm 2, ...\}$ a lexicographic order i.e., for $\mathbf{s} = (s_1,s_2)$ and $\mathbf{t} = (t_1,t_2) \in Z^2$, we write $\mathbf{s} \preceq \mathbf{t}$ if either $(s_1 < t_1)$ or $(s_1 = t_1$ and $s_2 < t_2)$.

We make the following assumptions:

i) $\{\varepsilon(i,j); (i,j)\epsilon Z^2\}$ is an independent and identically distributed $(i.i.d.)$ sequence of random variables with zero mean and variance σ^2.

ii) The $\beta_{kl}, k = 0, ..., p_1, l = 0, ..., p_2$, are real constants.

iii) $\{\beta_{kl}(i,j); (i,j)\epsilon Z^2\}$ is an independent and identically distributed $(i.i.d.)$ sequence of random variables with zero mean and variance w^2, and $(\beta_{kl}(i,j))_{0<k<p_1, 0<l<p_2}$ is independent of $(\varepsilon(i,j))$ for all $(i,j)\epsilon Z^2$.

The paper is organized as follows. In section 5, we give necessary and sufficient conditions for the existence of stationary and ergodic solution for equation (7). The condition ensuring the existence of moments of any order is given in section 6. Finally, covariance, spectral density for $SRCA(1)$ are obtained.

5 Stationarity

We use the following notation. Let $x = [X(i,j)]$ be an $(p_1+1)(p_2+1)$ matrix of observation on spatial series in plane.

We let $\underline{X}(i,j) = vec(x)$ denote the $P \times 1$ vectorization of the matrix x with $P = (p_1+1)(p_2+1)$. Then

$$\underline{X}(i,j) = \begin{bmatrix} X_i & X_{i-1} & \cdots & X_{i-p_1} \end{bmatrix}' \tag{8}$$

where $X_{i-k} = (X(i-k,j), X(i-k,j-1), ..., X(i-k,j-p_2))', k = 0, ..., p_1$. These imply that we can write the equations (7) in matrix form as

$$\underline{X}(i,j) = [W + \Pi(i,j)]\,\underline{X}(i,j) + \underline{\varepsilon}(i,j) \tag{9}$$

where $\mathrm{W}=\begin{bmatrix} W_0 & W_1 & \cdots & W_{p_1} \\ 0 & I_{p_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & I_{p_2} \end{bmatrix}, \Pi(i,j) = \begin{bmatrix} \Pi_0(i,j) & \Pi_1(i,j) & \cdots & \Pi_{p_1}(i,j) \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix},$

$\underline{\varepsilon}(i,j) = \begin{bmatrix} \varepsilon_i & 0 & \cdots & 0 \end{bmatrix}'$.

Note that $\underline{X}(i,j)$, $\underline{\varepsilon}(i,j)$ are $P \times 1$ dimensional vector, while W and $\Pi(i,j)$ are $P \times P$ dimensional square matrices with $E[\varepsilon_i \varepsilon_i'] = \zeta$, and $E[\Pi_k(i,j) \otimes \Pi_k(i,j)] = \Gamma_k$.

Our main result is contained that the spatial random coefficient autoregressive models (9) satisfies assumptions (i)-(iii) has a unique stationary and ergodic solution. This solution is given by

$$X_i = \sum_{n=0}^{\infty} T_i^n \varepsilon_{i-n}, \tag{10}$$

where the transition matrix T_i^n is defined as

$$T_i^n = \sum_{k=0}^{p_1} (W_k + \Pi_k) T_i^{n-k}, T_i^0 = I_{p_2}. \tag{11}$$

6 Existence of moments

In this section, we give the condition of finiteness of higher-order moments for the process $\underline{X}(i,j)$ defined by (9), which is satisfied the above assumptions.

We conclude by illustrate an example for $SRCA_2(1,1)$ process

$$X(i,j) = [\beta + \beta(i,j)]X(i-1,j-1) + \varepsilon(i,j). \tag{12}$$

we assume that the process $(X(i,j))_{(i,j)\in Z^2}$ is second order stationary, and we obtain its covariance function $C(h_1,h_2)$ and its spectral density f(λ).

Mot clés:
modèles autoregressifs à coefficients aléatoires spatials, statistiques spatiales, stationarité, densité spectrale.

Bibliographie

[1] Andel, J. (1991) On stationarity of multiple doubly stochastic model, *Kybernetika* vol. 27, No. 2.
[2] Aue, A. Horvath. L. and Steinebach. J. (2004) Estimation in random coefficient autoregressive models, *journal of time series analysis* vol. 27, No. 1.
[3] Choi, B. (1996) A recurcive algorithm for solving the spaial Yule-Walker equations of causal spatial AR models, *Stat- Prob letters* . 33, 241-251.
[4] Feigin, P. D. and Tweedie, R. L. (1985) Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments. *J. Times series analysis.*, vol. 6, No. 1.
[5] Guyon, X. (1995) Random fields on the network. *Springer-Verglag, New York.*
[6] Kashkovsky, D. V. and Konev, V. V. (2008) Sequential estimates of the parameters in a random coefficient autoregressive process. *Allerton Press Inc.*
[7] Major, P. (1981) Multiple Wiener-Itô integrals, In lecture Notes in Mathematics 849. *Springer-Verglag, New York.*
[8] Quinn, B. G. and Nicholls, D. E.(1981) Random coefficient autoregressive models: An introduction. (Lectures notes in statistics 11)*Springer-Verglas, Berlin-Heidelberg-New York.*

Statistical properties of Parasite Density estimators in Malaria and field applications

Imen Hammami*, André Garcia**, Grégory Nuel*

* Laboratoire MAP5 , 45 rue des Saints Pères, 75270 Paris Cedex 6, France
** Institut de Recherche pour le Développement IRD / UMR216, Laboratoire de Parasitologie, 4 avenue de l'Observatoire, 75270 Paris Cedex 6, France

ABSTRACT

Malaria is a global health problem responsible for nearly 3 million deaths each year, an average of one person every 12s. In addition, 300 to 500 million people contract the disease each year. The level of infection, expressed as the parasite density (PD), is classically defined as the number of asexual forms of *Plasmodium falciparum* relative to a microliter of blood. Microscopy of Giemsa-stained thick blood films is the gold standard for parasite enumeration in case of febrile episodes. PD estimation methods usually involve threshold values as the number of white blood cells (WBC) counted and the number high power fields (HPF) seen. However, the statistical properties of PD estimates generated by these methods have been generally overlooked.
Here, we study the statistical properties (bias, variance, False-Positive Rates ...) of the PD estimates of two commonly used threshold-based counting techniques according to varying threshold values. Furthermore, we give more insights on the behavior of measurement errors according to varying threshold values and on what would be the optimal threshold values that minimize the variability.

Keywords: Threshold-based counting techniques, parasite density estimators, bias, variance, False-Positive Rates, malaria epidemiology.

RÉSUMÉ

Le paludisme est un problème mondial de santé publique. Cette maladie provoque environ 3 millions de décès par an. On recense 300 à 500 millions de cas clinique chaque année. La charge parasitaire est le critère essentiel du diagnostic d'accès palustre devant un syndrome fébrile. La densité parasitaire (DP) est exprimé en nombre de *Plasmodium falciparum* au stade de développement asexué par microlitre de sang. L'examen de la goutte épaisse est la technique de référence pour l'estimation de la DP dans les études portant sur le paludisme. L'estimation repose souvent sur l'utilisation de techniques de comptage à seuils. Ces seuils représentent le nombre de leucocytes comptés et/ou le nombre de champs microscopiques lus.

Les propriétés statistiques des estimateurs de la DP résultant d'approche de ce type n'ont pas été traité dans la littérature. Dans cet article, on s'intéresse à l'évaluation de deux méthodes d'estimation à travers l'étude des propriétés statistiques des estimateurs de la DP (le biais, la variance, le taux de faux positifs ...) en fonction de seuils variables. On montre qu'un choix approprié des valeurs seuils permet de minimiser la variabilité statistique des méthodes d'estimation en fonction des niveaux de qualité recherchés.

Mot-clès: Méthodes d'estimation à seuils, estimateurs de la densité parasitaire, biais, variance, taux de faux positifs, coût, épidémiologie du paludisme.

Introduction

Accurate estimation of the PD is an important endpoint in epidemiological studies and clinical trials, both as a direct measure of the level of infection in a population and when defining parasitemia thresholds to diagnose malaria in case of fever episodes. Malaria PD estimates are also used to assess the development of naturally acquired immunity [6] and in malaria vaccines investigations. Therefore, inaccurate estimation of PD can lead to patient mismanagement and public health misinformation.
The usual and most reliable diagnosis of malaria is microscopic examination of thick films. The thick films are used to detect infection, and to estimate parasite concentration or parasite density (PD). The level of infection, expressed as the parasite density is classically defined as the number of asexual forms of plasmodium per a fixed volume of blood (e.g. microliter). In the following, the term of parasite refers to asexual form of Plasmodium.
Parasite density estimation methods usually involve threshold-based counting techniques. Threshold definition and values may vary a lot from one method to another. Parasites may be counted either in relation to the number of microscopic high power fields (HPF) [1], or according to the number of leukocytes (WBC) which are simultaneously counted. In the first case the methodology would be as follows: if less than n parasites were counted in the m first HPFs, then do this, else do that. In the second one, the m first HPFs will be substituted by "when N leukocytes are counted" . Conversion to counts per microliter then depends on an assumption of the volume of blood examined.
To deal with potential inaccurate estimations of the PD, research teams aim to analyze more slides and subjects, to improve repeatability and to decrease the variability. However, one of the problems they have to deal with is that during large scale studies the number of thick blood smears performed can be largely greater than 10000. Then, the microscope slides rereadings lead to an important cost overrun in term of both money and time. Out of the problem mentioned previously, a important question arises here on what would be the real interest these practices may have on the final results. In this paper, we try to give insights on how to handle this inescapable generated variability with an appropriate choice of threshold values instead of systematic rereading procedures.

[1] defined as oil immersion microscopic field x 1000

Materials & Methods

Threshold-based counting techniques

There are several threshold-based counting techniques commonly used in epidemiological surveys. In this paper, we are interested in methods where parasites and leukocytes are counted simultaneously. *Trape* (1985) [8] proposed to count the total number of parasites per 200 white blood cells (`WBC`) and to multiply this number by 40 to give the number of parasites per μl. This average value of 8000 `WBC` per μl is accepted as reasonably accurate by `WHO` [2] [5]. Different counting methods have been proposed by research teams in order to optimize the assessment of `PD`. Among them, *Garcia et al.* (2004) [4] proposed to count the parasites until either 500 `WBC` or 500 parasites have been seen in an epidemilogical survey conducted in the *Tori Bossito* area (*Southern Benin*). In the following this method will be referred as *Tori Bossito* method.

Variability

The importance of parasite density data reproducibility stems from the need for epidemiological interpretations to be based on solid evidence. However, variation of parasite density within a slide is expected even when prepared from a homogeneous sample [1]. The source and scale of measurement error (sample preparation, staining process, counting technique, microscopist performance) have been investigated. Inter-rater reliability is a source of concern in this context. It refers to a metric for raters consistency that measures the degree of agreement among raters [3]. Many techniques were developed to measure inter-rater reliability. A number of reports attempted to evaluate the inter-rater reliability of malaria microscopy in epidemiological surveys and to quantify the degree of agreement between malaria slide density readings through statistical approaches [1]. For continuous data, Analysis of Variance (`ANOVA`) is the method of choice. *Bland & Altman* (1986)[2] plotted the differences in log-transformed data versus average in mean counts. They have expanded on this idea by graphing the difference of each point, the mean difference, and the confidence limits on the vertical against the average of the two ratings on the horizontal. The resulting *Bland & Altman* plot demonstrates not only the overall degree of agreement, but also whether the agreement is related to the underlying value of the item. For instance, two raters might agree closely in estimating the size of small items, but disagree about larger items. *Alexander et al.* (2010) [1] assess agreement between replicate slide readings of malaria parasite density using as criteria the the repeatability (r) which is the value bellow which the absolute difference between to results may be expected to lie with the probability of 95%. This metric is linked to *Bland & Altman* [2] limits of agreement. It's half the distance between the upper and lower limits of agreements. For nominal data, the *kappa* (κ) [3] coefficient of *Cohen* and its many variants and the *Scott*'s *pi* (π) [7] are the preferred statistics.

[2] World Health Organisation

[3] A rater refers to any data-generating system, which includes microscopists and laboratories.

Along the same lines as the experimental variability, the sampling variability is a source of interest while studying the efficiency of estimation methods. Sampling variability refers to the different values which a given function of the data takes when it is computed for two or more samples drawn from the same population. Sampling errors and biases are induced by the sample design. They include the selection bias when the true selection probabilities differ from those assumed in calculating the results and random sampling error which is the random variation in the results due to the elements in the sample being selected at random. A measure of the sampling error is the standard error. This measurement is based on the idea of selecting several samples. Sampling variability can also be expressed relative to the estimate itself through the coefficient of variation (`CV`) [4].

Though, through a literature review, we noticed that none of the studies of variability dealt with the sampling error generated by the threshold-based counting techniques or evaluated the impact of the threshold values in endpoint measurements. In addition, the accuracy and consistency of these methods have been generally overlooked. Furthermore, there is no general agreement on the optimal method for estimating the parasite density according to thresholds values. Further experimental evidence is needed to determine which parasite counting technique is most accurate, reproducible, and efficient. However, the homogeneity question remains. The distribution of the thickness of the smear and hence the distribution of parasites within the smear is not completely homogeneous. Therefore, a proportion of the variability may be explained by this homogeneity factor.

Methodology

In the following, we discuss two methods of parasites density estimation. The first method we focussed on is counting the parasites until 200 `WBC` have been seen. The second one is counting the parasites until either 500 `WBC` or 500 parasites have been seen. We will discuss these two methods in terms of bias, variance, truncation levels with fixed threshold values. Furthermore, we show that we can handle the variability by juggling with threshold values. In the following, we generate recursive density estimates under homogeneity and uniformity assumptions of the parasite density distribution throughout the thick film. Simulation models assumptions were validated throught proper statistical tests. We considered the problem of testing whether the parasite counts per `HPF` comes from a Poisson distribution. Our interest is the alternative that the observations are overdispersed with respect to a *Poisson* distribution, for which the mean is equal to the variance. Hence a *Poisson* distribution is not an appropriate model. The alternative models are the negative binomial distribution and the zero-inflated *Poisson* models. Furthermore, we tested for heterogeneity by computing the log likelihood differences between the heterogeneous model (*Poisson* mixture model) and the homogeneous model (single *Poisson* model).

[4] CV is defined as the ratio of the standard error to the mean.

Results

Different threshold values may be fixed, which raises questions regarding accuracy and reproducibility. A key question is to which extent the threshold values would influence the variability in parasite density estimates. To understand how the thresholds involved in parasite enumeration methods contribute to the magnitude of discrepancies in PD determination, we measure the bias and variance and False Positive Rates of two threshold-based counting techniques.
All the statistical properties of the PD estimators were expressed as functions of the PD and the threshold values. For brevity, we show only a part of our results.

In the following, we will be interested in the *Trape* method. We compute the probability density function of the PD estimator. We derive the expression of the bias as well as the standard deviation which are expressed as a percentage of the PD. We calculate the truncation levels generated by the *Trape* method and we assess the method cost in terms of number of HPF read. We prove that the relative bias depends only on thresholds values (density independant). The figure below shows that the relative bias decreases as threshold values increase.

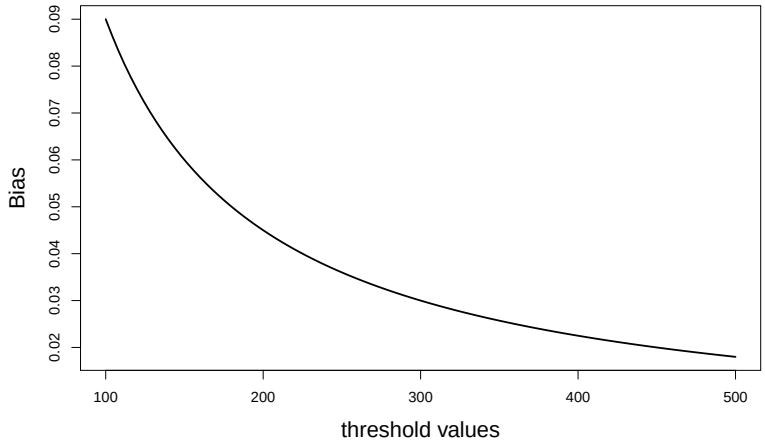


Figure 1: Relative bias of the *Trape* counting method according to threshold values.
For instance, counting parasites until 500 WBC instead of 300 WBC decreases noise by 1%. Thus far this has resulted in up to 1.67-fold difference in costs. This can help to decrease any bias that we might see in the samples with an appropriate choice of the WBC threshold value.

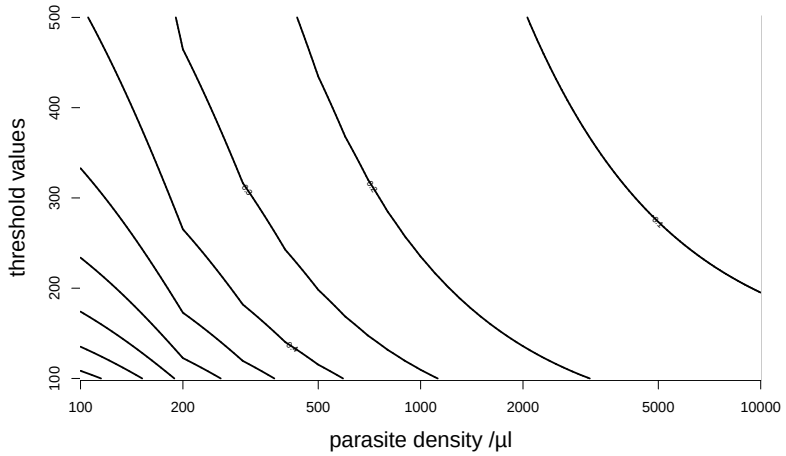


Figure 2: Relative standard deviation (*std*) of the *Trape* method according to parasite densities (per μl) and threshold values. The color map draw attributes define a gradient of colors related to generating regions of *relative std* values as the first color of the domain refers to the minimum value. The contour lines define constant levels of *relative std*. For high signals, the *relative std* lies midway between 10% and 20% of the PD. For low signals, *relative std* is up to 30%.

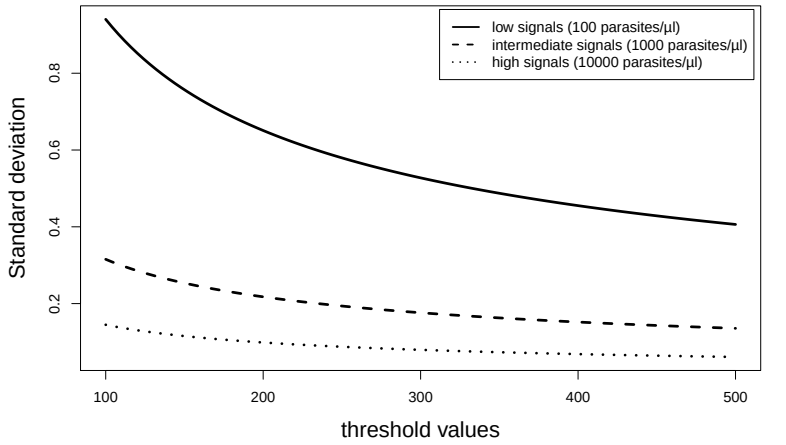


Figure 3: Relative standard deviation cross-cuts for low, intermediate and high signals.
To understand how the threshold values have an influence on the variability of the *Trape* counting method estimates, we plotted the relative *std* according to threshold values for 3 fixed parasite densities. We see very few variation of the relative *std* as threshold values increase.

References

[1] N. Alexander, D. Schellenberg, B. Ngasala, M. Petzold, C. Drakeley, and C. Sutherland. Assessing agreement between malaria slide density readings. *Malaria Journal*, 9(1):4, 2010.

[2] M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1:307–310, 1986.

[3] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

[4] A. Garcia, A. Dieng, F. Rouget, F. Migot-Nabias, J. Le Hesran, and O. Gaye. Role of environment and behaviour in familial resemblances of plasmodium falciparum infection in a population of senegalese children. *Microbes and Infections*, 6:68–75, 2004.

[5] W. H. Organization. *Basic Malaria Microscopy: Part I. Learner's Guide, Second Edition*, April 2010.

[6] C. Rogier and J. F. Trape. Malaria attacks in children exposed to high transmission: who is protected? *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 87:245–246, 1993.

[7] W. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 17:321–325, 1955.

[8] J. F. Trape. Rapid evaluation of malaria parasite density and standardization of thick smear examination for epidemiological investigations. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 79:181–184, 1985.

Méthodologie d'inversion des mesures optiques (AOT) en mesures de qualité de l'air (PM10) basée sur les réseaux de neurones.

H.Yahi [1,2], S.Thiria[1], M.Crepon[1], A.Weill[3], R.Santer[2]
[1] IPSL/LOCEAN, UPMC/CNRS/IRD/MNHN, Universitế Pierre et Marie Curie, 4 place Jussieu, 75252 Paris Cedex 05, France
[2] Université du Littoral côte d'opale ULCO, F-62930bWimereux, France
[3] IPSL/LATMOS, Université de Versailles Saint Quentin en Yvelines, UMR 8190 Guyancourt (78), France

Résumé
Nous présentons une méthode d'inversion de mesures photométriques d'épaisseurs optiques (AOT) en concentrations massiques atmosphériques (mesures de qualité de l'air PM10). Comme les PM10 est une mesures de surface et l'AOT est une mesure intégrée, la relation liant les deux mesures est très complexe. Étant donné que ces deux paramètres dépendent fortement de structures atmosphériques et des paramètres météorologiques, nous avons classé les situations météorologiques en termes de types de temps en utilisant un classificateur neuronal (cartes auto-organisatrices). Pour chaque type de temps, nous avons constaté que la relation entre l'AOT et de PM10 peut être établie et l'inversion effectuée à l'aide de réseaux de neurones avec des performances satisfaisantes. Afin d'accroître la fiabilité statistique de la méthode nous avons appliqué cette approche à la région de Lille (France) pour la période de cinq d'été (étés des années 2003-2007).

Abstract
We present a method for retrieving atmospheric particulate matter (PM10) from sun-sky photometer measurements (AOT). As PM10 is a "surface parameter" and AOT is an "integrated parameters", we first determined whether a "functional relationship" linking these two quantities exists. Since these two parameters strongly depend on atmospheric structures and meteorological variables, we classified the meteorological situations in terms of weather types by using a neuronal classifier (Self organizing Map). For each weather type, we found that a relationship between AOT and PM10 can be established. We applied this approach to the Lille region (France) for the summer 2007 and then extended to a five summer period (summers of the years 2003e2007) in order to increase the statistical confidence of the PM10 retrieval from AOT measurements. The good performances of the method led us to envisage the possibility of deriving the PM10 from satellite observations.

Mots clés : Exploitation des données, cartes topologiques, classification, apprentissage non supervisé, problèmes inverses, environnement, pollution.

Introduction
Les aérosols constituent un enjeu important pour l'étude de l'atmosphère. En effet ceux-ci ont des conséquences sanitaires néfastes, modifient-la visibilité et ont sur le bilan radiatif des effets directs (absorption/ réfraction de la lumière), semi directs (modification du profil vertical de température) et indirects (impacts sur la formation des nuages en tant que noyaux de condensation). Prévoir la concentration de la pollution en aérosols (PM10) est donc important.
Les mesures optiques d'épaisseur optique en aérosols (AOT) permettent la surveillance globale journalière du contenu atmosphérique en particules. Cependant la relation permettant de retrouver les concentrations massiques en (PM_{10}) à partir des mesures optiques (AOT) est loin d'être simple.
Malgré de nombreuses méthodes habituellement utilisées pour l'estimation de la concentration des aérosols aucune d'entre elles n'est communément admise et ne donne des résultats satisfaisants. Les méthodes statistiques sont devenues de plus en plus efficaces pour extraire de l'information à partir des données disponibles.

Cet article présente une méthodologie d'inversion de l'épaisseur optique atmosphérique (AOT) en concentration massique en aérosols (PM10) dans les zones urbaines, basée sur des méthodes statistiques d'exploitation de données utilisant les réseaux de neurones.
Cette méthodologie est basée sur une classification en types de temps qui permet de mettre en évidence des situations météorologiques typiques à partir de la classification des paramètres météorologiques modèle (Vautard (1990); Bissolli et Dittmann (2001) ; Noordijk et Visser (2002)) pour lesquelles, la relation (PM_{10}, AOT) est simplifiée, et ensuite une inversion d'AOT en (PM10) pour chaque type de temps avec des performance largement meilleure. Les méthodes statistiques utilisées sont les cartes auto-organisatrices SOM (Self Organizing Map, Kohonen(1982)) qui sont des méthodes neuronales de classification, et la classification hiérarchique ascendante (Jain et Dubes (1988)).

Données d'entrée du modèle

Nous avons réalisé une étude d'inversion des AOT en mesure de qualité de l'air au sol dans la région de Lille pendant les cinq étés successifs (2003 à 2007). Nous disposions de mesures de qualité de l'air (PM10) faites à partir de stations au sol et de mesures d'épaisseur optique (AOT) à trois longueurs d'onde faites à partir d'un photomètre.
Au moment où le modèle a été réalisé, les données d'entrée suivantes étaient disponibles:

- les profils verticaux (jusqu'à une altitude de 12 km) des paramètres météorologiques de vent zonal, vent méridien, température, fournit par le modèle météorologique MM5 avec une résolution spatiale de 0.5° et temporelle d'une heure.
- la pollution au sol mesurée chaque heure à partir les stations de surveillance automatique régional de qualité de l'air dans l'agglomération Lilloise.
- Les mesures horaires d'épaisseur optique atmosphérique (AOT) dans le visible (c'est à dire entre 5h à 18h) en utilisant une station de mesure au sol appelé photomètre dans l'agglomération Lilloise.

Du point de vue de la disponibilité des données, les sorties modèle MM5 de profils verticaux des paramètres météorologiques de vent et de température ainsi que les mesures in-situ de la qualité de l'air sont disponibles toutes les heures en routine, et entre 5h et 18h, à condition d'absence de nuages pour les mesures d'épaisseur optique atmosphérique (AOT).
Sur les cinq étés d'étude de la région Lilloise, nous avons compté 9917 mesures horaires de qualité de l'air PM_{10}, de profils de paramètres météorologiques (vent et température), et seulement 836 mesures horaires d'épaisseur optique AOT aux trois longueurs d'onde (430nm, 760nm, 870nm). Le peu de mesures d'épaisseur optique par rapport aux autres données est due à la présence de nuages.

Méthodologie

Il est bien connu que la survenu des fortes pollutions est associée aux conditions météorologiques, ce qui nous a incité à regrouper les situations similaires correspondants aux mêmes niveaux de pollution de l'air.
Le modèle d'inversion de l'épaisseur optique (AOT) pour retrouver les concentrations massiques (PM10) se fait en deux étapes (Yahi et al (2011)): classification des facteurs météorologiques en types de temps en utilisant seulement les paramètres météorologiques, ensuite l'inversion de l'épaisseur optique atmosphérique (AOT) en concentration massique (PM10) pour chaque type de temps.
La classification des facteurs météorologiques en types de temps est réalisée à partir des sorties d'un modèle météorologique semi-opérationnel, (le modèle MM5 forcé aux frontière par le modèle opérationnel ECMWF). Pour cela on a utilisé les cartes topologiques (algorithme SOM) suivi d'une classification hiérarchique (CAH). Chaque vecteur d'entré de l'algorithme SOM (Self Organizing Map) est composée du profil vertical de vent zonal, vent méridionale et de température, sur chacune des 12 mailles sélectionnées autour de la région Lilloise (Figure 1), cela pour donner aux types de temps une structures météorologiques méso-échelle. Il est donc de dimension (720 =12*(3*20))

représentant les 12 mailles prises en considération pour caractériser les types de temps sur la région de Lille et en chaque maille les données MM5 (les trois variables météorologique) sur les 20 niveaux (constituant chaque profil).

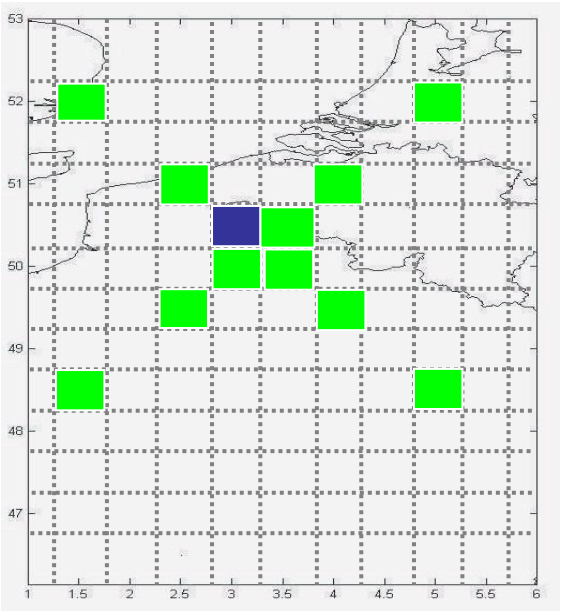


Figure 1- Représentation de la zone d'étude et des mailles pour lesquelles on considère les données de vent méridional, zonal, température produites par MM5. La maille bleue représente l'agglomération Lilloise.

Étant donnée la dimension de la base de données étudiée, nous avons choisi une carte SOM carrée de grande dimension (10*10 référents). Cette carte permet de faire une répartition de la base de données météorologique en 100 groupements de situations météorologiques similaires. Chaque référent représente donc une situation météorologique typique. Nous avons rassemblé ensuite les référents en groupes en appliquant une classification hiérarchique (CAH) à la carte auto-organisatrice. L'utilisation de la CAH est basée sur un critère de similarité ; selon ce critère, on agrège les deux groupes les plus proches pour former un nouveau groupe. Ce nouveau groupe peut, à nouveau être agrégé. On procède ainsi de façon itérative jusqu'à n'avoir plus qu'un seul groupe. Cette méthode produit un arbre qui résume la succession de ces regroupements appelé dendrogramme. Le niveau auquel ce dendrogramme est coupé permet de déterminer le nombre de classes (ou clusters) produites.

La classification retenue est celle issue d'une classification ascendante hiérarchique en huit classes avec le critère de Ward comme critère de similarité et une contrainte de voisinage.

Nous avons obtenu donc huit types de temps, dont chacun correspond à un niveau de pollution, la Figure 2, résume la première étape de la méthodologie.

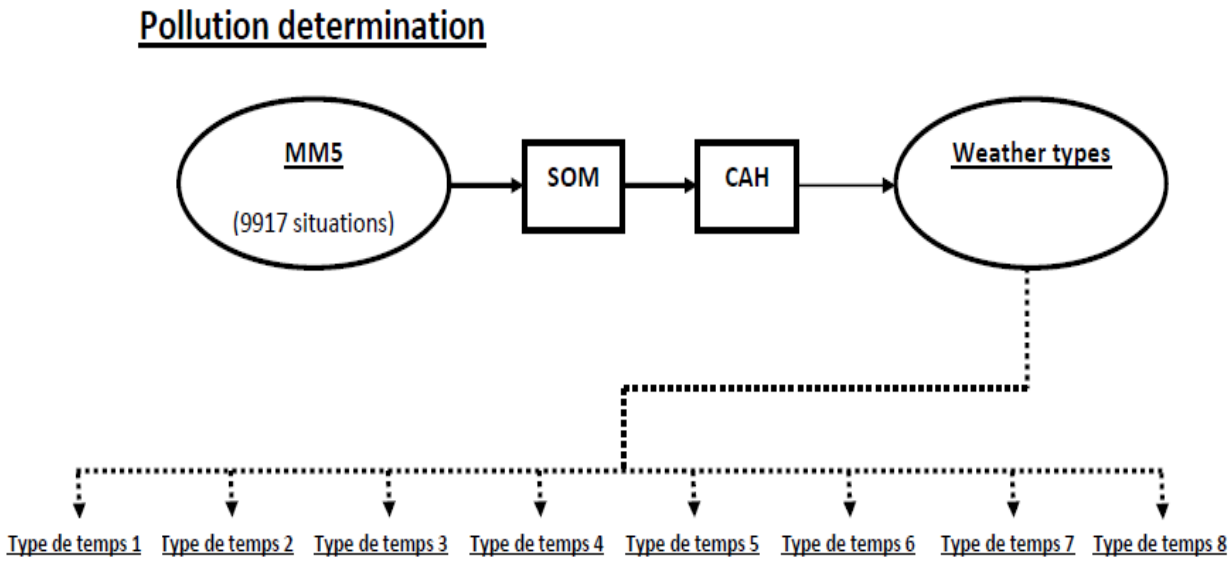


Figure 2 – Schéma fondamental de la méthode ; le haut de la figure représente la classification des situations météorologiques en classe ; le bas montre l'association entre les classes et les types de temps.

Les 836 situations météorologiques sur Lille associées aux 836 mesures d'AOT, ont été projetées sur la carte SOM et réparties entre les huit classes. Le Tableau 1, montre le nombre de mesures d'AOT et de PM10 qui vont constituer l'ensemble d'apprentissage (80%) et de test (20%) pour effectuer l'inversion (AOT→ PM10) dans chaque classe de type de temps.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8
Ensemble d'apprentissage	**56**	**105**	**25**	**66**	**53**	**99**	**206**	**66**
Ensemble de test	**27**	**25**	**4**	**13**	**12**	**11**	**49**	**19**

Tableau 1 - Nombre de données participant à l'apprentissage et au test pour chaque classe.

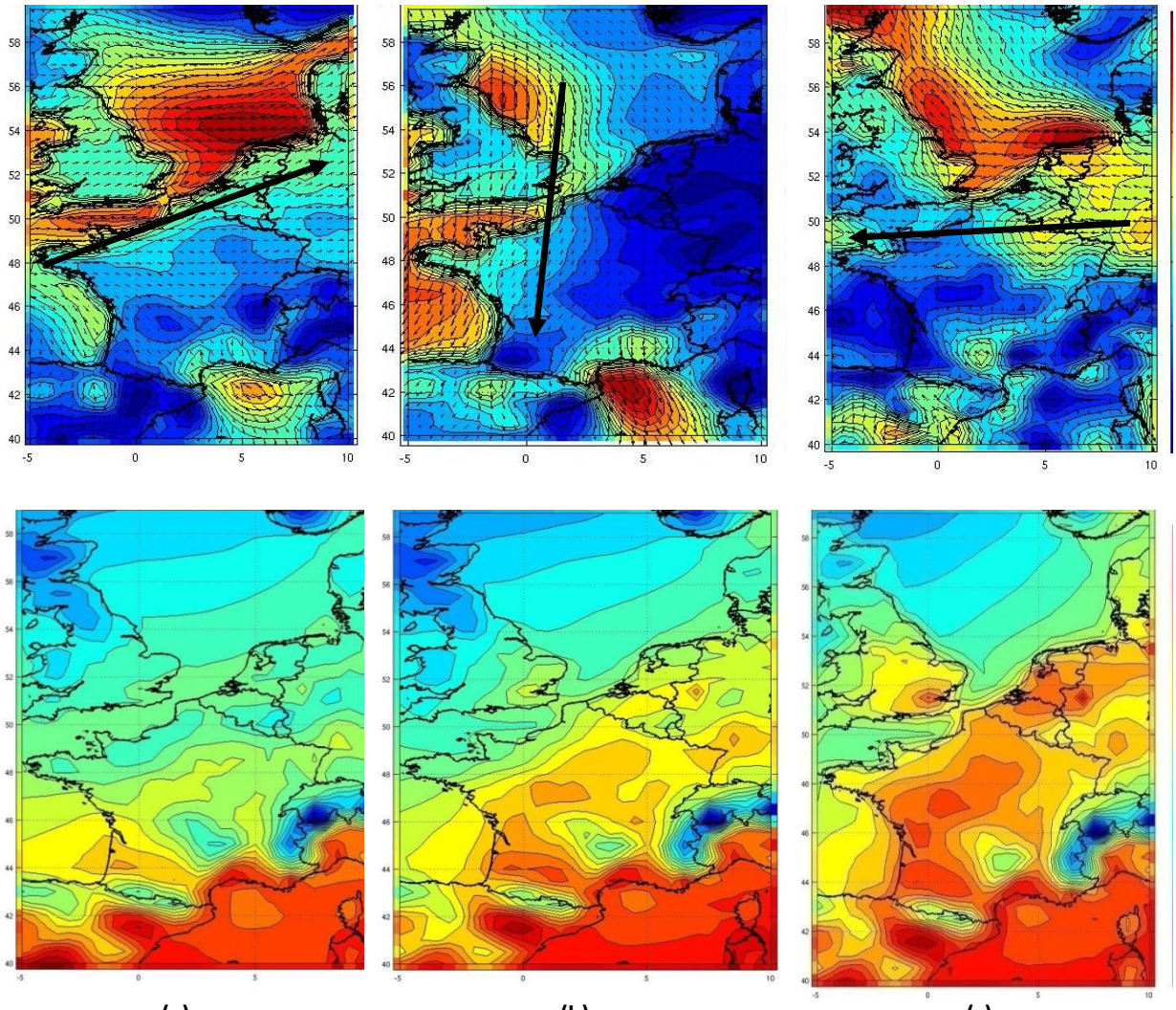
Pour caractériser chaque classe de type de temps par un indice de pollution, nous avons effectué la moyenne des valeurs d'AOT et de PM10 de l'ensemble d'apprentissage de chaque classe. Ces moyennes représentent une estimation de l'état de la pollution des classes. Nous proposons d'utiliser les valeurs moyennes d'AOT comme indice de pollution:
Indice 1 : AOT<0.11, ce qui correspond à une pollution faible (classe1).
Indice 2 : 0.11≤ AOT< 0.17, ce qui correspond à une pollution moyenne ou habituelle (classe 2 et 3).
Indice 3 : 0.17≤ AOT< 0.23, ce qui correspond à une pollution forte (classe 4,5 et 6).
Indice 4 : 0.23≤ AOT, ce qui correspond à une pollution très forte (classe 7 et 8).

Nous allons aussi utiliser les cartes auto-organisatrices pour l'inversion de l'épaisseur optique atmosphérique AOT en concentration massique en aérosols PM10 et cela pour chaque classe. Afin d'affiner la relation AOT→PM_{10} que nous voulons déterminer, nous utiliserons comme variables d'entrée de la nouvelle carte auto organisatrice SOM2, les spectres d'AOT et les variables météorologiques locales (profil de vent zonal, vent méridionale, température à Lille). On a donc huit cartes SOM2. Chaque neurone de SOM2 peut alors être associé à une moyenne de PM_{10}. De cette manière, on obtient huit fonctions non paramétriques permettant de prédire les relations AOT →PM_{10}. Pour plus de précision dans l'inversion, nous avons choisi une carte SOM2 carrée de 36 neurones (6*6). Pour chacune des huit cartes apprises, le Tableau 1, montre le nombre de données

utilisées pour l'apprentissage et le test pour chaque classe. Bien entendu, un nombre plus grand d'observations, doit permettre d'affiner encore les relations que nous allons établir.

Résultats de l'inversion

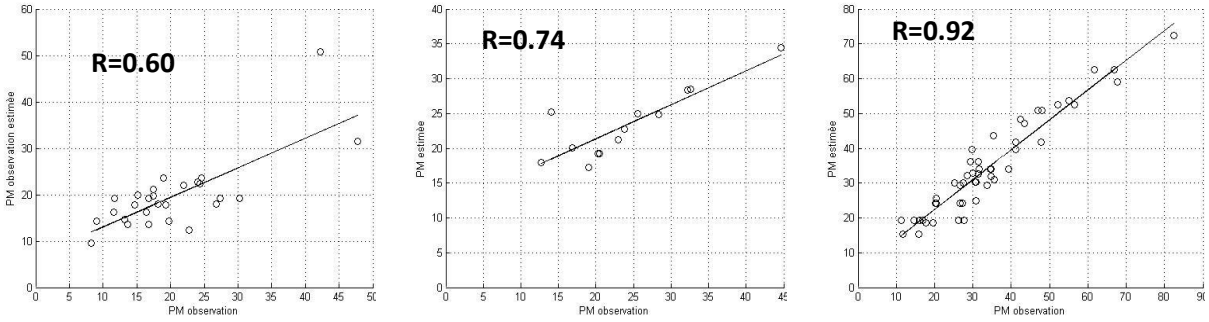
Nous avons obtenu huit types de temps, chacun caractérisé par une direction du vent et une température moyenne sur la région Lilloise. Les huit types de temps sont très différents les uns des autres ce qui montre la pertinence de cette classification. Chaque type de temps a été caractérisé aussi par un indice de pollution. A titre d'exemple, nous montrons sur la Figure 3, trois types de temps sur les huit obtenus. Nous avons choisis de monter trois types de temps caractérisés par des indices de pollution différents (allant du faible au fort).



(a) **(b)** **(c)**

Figure 3 : cartes moyennes de vent et de température obtenues sur la France pour : (a)-classe2, (b)-classe4 et (c)-classe7.

Les performances du modèle SOM2 pour l'inversion sont très satisfaisantes surtout pour les types de temps caractérisés par des indices de pollution forts (coefficient de corrélation, R=0.92), ce qui est remarquable vu l'importance de la prévision des forts épisodes de pollution.
La Figure 4, montre les diagrammes de dispersion entre les mesures de qualité de l'air et les PM10 estimées à partir de l'inversion de l'AOT en utilisant notre modèle d'inversion pour les trois types de temps montrés sur la Figure 3.



(a) **(b)** **(c)**

Figure.4 – Diagramme de dispersion entre les mesures de qualité de l'air et les estimations des concentrations massiques obtenu par inversion de l'AOT pour les classes : (a)-classe 2, (b)-classe 4 et (c)-classe 7.

Conclusion

La méthodologie proposée est une approche originale de la prévision de la pollution de l'air à partir des AOT. Elle présente une alternative aux méthodes utilisées jusqu'à présent. Les résultats obtenus sont très encourageants. Tout d'abord, pour les indices de pollution, nous avons montré que notre méthodologie fournit un indice de pollution à partir de la prévision météorologique.

Cette méthodologie a été validé aussi pour l'inversion des AOT en PM10 pour les huit classes avec un nombre de situations statistiquement significatif et les performances d'inversion sont très satisfaisantes surtout pour les classes de forte pollution (R=0.92).

Bibliographie

[1] Bissolli et Dittmann (2001) The objective weather type classification of the German weather service and its possibilities of application to environmental and meteorological investigations, METTOOLS conference No 4 Stuttgart, 10, no4, pp. 227-281 (1 p.1/4), pp. 253e260.

[2] Kohonen(1982) Self-organized formation of topologically correct feature maps, Biol. Cybern. 43, 59e69.

[3] Jain et Dubes (1988) ***Algorithms for Clustering Data*****. Prentice-Hall advanced reference series, Prentice-Hall, Inc., Upper**

[4] Noordijk et Visser (2002) Correcting air pollution concentrations for meteorological conditions; an extended regression-tree approach, Adopted from the Netherlands Environmental Balance, 2002 & 2003 editions.

[5] Vautard (1990) Multiple weather regimes over the north Atlantic: analysis of precursors and successors, Mon. Wea. Rev. 118, 2056e2081.

[6] Yahi et al (2011) Exploratory study for estimating atmospheric low level particle pollution based on vertical integrated optical measurements, Atmospheric Environment.

Modélisation Multi-variée des extrêmes hydrométéorologiques

Application: Mildiou de la vigne

Auteurs :

Dhouha OUALI : Etudiante chercheur à l'ENIT ; BP 37, Le Belvédère 1002 Tunis.

Zoubeida BARGAOUI : Professeur à l'ENIT ; BP 37, Le Belvédère 1002 Tunis.

Samir CHBIL : Docteur et chercheur en phytopathologie, Technopôle de Borj Cedria -BP 95 Hammam Lif 2050.

Résumé :

La notion du risque est une notion très complexe ; elle découle de la conjonction d'un aléa non maîtrisé et de l'existence d'un enjeu ou d'un environnement pouvant être affecté par un tel événement. Ainsi, nous visons par la présente étude développer un processus d'identification du risque, en se basant sur la théorie statistique des valeurs extrêmes. En fait, cette dernière présente un outil adéquat de maitrise et d'aide à la décision via la modélisation de l'occurrence des événements extrêmes. Notre étude consiste à étudier le risque lié au Mildiou de la vigne dans les vignobles du Cap Bon-Tunisie, et à proposer une approche de précaution afin d'obtenir les objectifs fixés en termes de qualité et de quantité avec un minimum d'interventions.

Les variables à modéliser sont la température et les précipitations ; du fait de la structure de corrélation de ces deux variables, une analyse fréquentielle univariée ne permettra pas de bien évaluer les probabilités au dépassement, ainsi nous avons eu recours à la théorie des copules pour effectuer une modélisation bivariée de ces deux variables.

Abstract:

The notion of risk is a very complex concept; it results from the conjunction of an uncontrolled random and the existence of an issue or an environment that may be affected by such an event. Thus, we try in this study to develop a process of risk identification, referring to the statistical theory of extreme values. In fact, the latter is an appropriate tool to make a decision through modeling the occurrence of extreme events. Our study consists on studying the risk associated to the Mildiou of grapes in the vineyards of the Cape Bon Tunisia, and to propose a precautionary approach to achieve the objectives set in terms of quality and quantity with a minimum of interventions.
Variables to model are temperature and precipitation; Because of the correlation of these two variables, a univariate frequency analysis will not accurately assess the probability of such

excess, so we resorted to the theory of copula to perform bivariate modeling of these two variables.

Mots-clés : Mildiou, risque, copules, corrélation, quantiles, bivarié, bootstrap, distribution jointe.

1. Introduction

Du fait des caractéristiques de la zone d'étude (Cap bon-Kélibia), nous sommes intéressés dans ce travail à étudier la vulnérabilité du secteur agricole face aux agents climatiques. En fait, le gouvernorat de Nabeul seul participe à 16% de la production agricole nationale ; en outre, la région produit environ 80 % du raisin tunisien, cultivé sur une surface agricole d'environ13 500 hectares (soit les deux-tiers de la surface totale destinée à la production du vignoble). De ce fait, nous nous sommes rendu compte de l'importance des précautions qu'il faut adopter pour minimiser les dégâts des risques qui peuvent toucher à ce secteur.

2. Mildiou de la vigne

Un des risques les plus ravageurs pour la viticulture est celui des maladies cryptogamiques telles que l'Oïdium, la pourriture grise et le Mildiou. Toutefois, ce dernier est le plus dommageable, et est à risque occasionnel du fait qu'il dépend énormément des conditions météorologiques ; les pertes de rendement enregistrées dans les parcelles contaminées sont de plus de 30 %, voire beaucoup plus dans les parcelles les plus touchées. Il s'agit en fait, d'un champignon microscopique à mode de vie aquatique ; il passe par une phase active, d'avril à octobre, et passe l'hiver à l'état de repos sous formes d'oospores à l'intérieur des feuilles mortes tombées au sol. Une atmosphère humide, un temps pluvieux et une température douce constituent les conditions les plus favorables au développement du Mildiou [5].

Il est à noter que du fait da la nature de cette épidémie, il est possible d'éviter les dégâts énormes qui peuvent en résulter, en adoptant une stratégie de lutte préventive. Ceci se résume à appliquer les recommandations suivantes :

- Soigner le choix des variétés lors de l'achat: choisir celles annoncées comme peu sensibles ;
- Assurer une bonne fertilisation, ce qui rend les plantes plus résistantes ;
- Tailler les rameaux et les attacher correctement, pour favoriser une bonne aération et un séchage rapide ;
- Ramasser et brûler soigneusement et régulièrement les feuilles tombées au sol à l'automne.

3. Méthodologie

L'objet de cette étude s'intègre dans le cadre d'une stratégie de gestion du risque; en effet, nous cherchons à modéliser la probabilité d'occurrence du Mildiou en précisant les conditions favorables à son déclenchement, ainsi que de prévoir ses périodes de retour.

En fait, la théorie du risque est largement utilisée dans de nombreux secteurs d'activité (assurance, économie, …) notamment l'épidémiologie, dont les étapes consiste à : identifier les facteurs du risque, évaluer les risques, les maitriser.

Or, pour expliquer une telle approche, nous sommes amenés à définir sa terminologie ; on appelle risque la probabilité de survenue d'un événement ayant des impacts indésirables sur un milieu particulier. Toutefois, selon Stirling, lors d'un traitement d'un problème scientifique donné, nous sommes face à plusieurs autres aspects à savoir l'ambigüité, l'incertitude et l'ignorance [1].Le tableau suivant détaille les spécificités de chacun de ces termes en ce qui concerne notre cas d'étude.

Tab1 : Identification des facteurs : risques, Ambigüité, incertitude et ignorance selon notre cas d'étude

Risque : • Evaluation du risque : - Les facteurs de risque : précipitation élevée, température douce, humidité élevée - Exposition au risque : La contamination de la vigne débute vers le mois d'Avril, Mars ou Juin. - Modélisation statistique des variables climatiques.	**Ambigüité** : • Aspects : - La description qualitative de l'épidémie : La terminologie utilisée
Incertitude : • Les facteurs : - La fiabilité des données - L'éloignement du site - La représentativité de l'échantillon - Les ajustements des lois.	**Ignorance** : • Suivi et surveillance : - Les données relatives au suivi du Mildiou (les années d'occurrence de l'épidémie).

Afin de réduire le niveau du risque, nous développons dans notre étude les étapes relatives à un processus d'évaluation du risque en procédant par une analyse statistique des facteurs climatiques (Température et précipitations). Cette analyse consiste à calculer les probabilités jointes de dépassement .Entre outre, on utilise la théorie des copules pour mieux décrire la structure de dépendance entre les deux variables climatiques et par suite reproduire convenablement les périodes de retour [4]. Pour ce faire, une étude univariée des lois marginales des variables doit être réalisée ; ensuite, une méthode de ré-échantillonnage (Bootstrap) est introduite pour étudier les incertitudes sur l'estimation des paramètres des lois d'ajustements marginales. Après la détermination de la meilleure copule à adopter, nous seront aptes de calculer les quantiles bivariés au dépassement [3], et par suite de déterminer les périodes de retour bivariées.

Entre autre, nous entamons une étude de sensibilité de nos résultats aux seuils de déclenchement du stade active de l'épidémie (un seuil égale à 1 mm, 5 mm…).

4. Résultats et perspectives

La figure 1 représente les valeurs de la température moyenne et du cumul de pluie de la saison définit sur les mois Avril-Mai-Juin, durant toute la période de mesure (1970-2000). On observe une légère dépendance négative entre les différentes paires de variables.

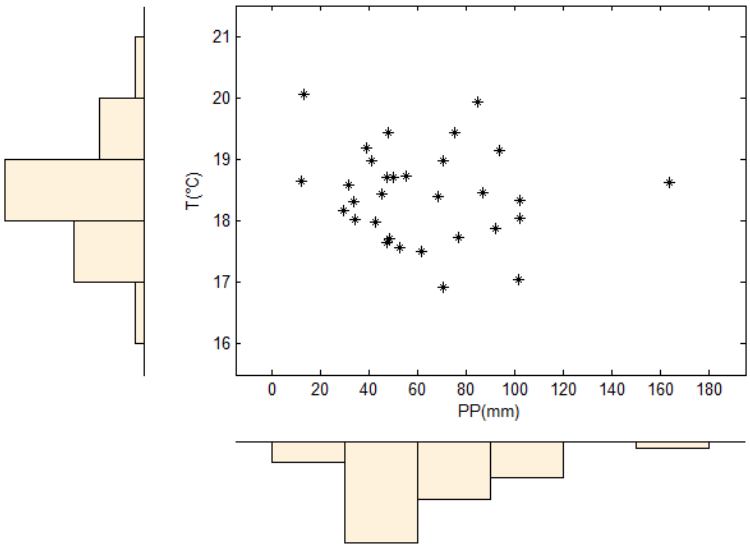


Fig1 : Graphique des paires (Température, Précipitations) et densités marginales associées

La figure 2 illustre le résultat d'un ajustement de la variable Précipitation à la loi Log-Normal; les paramètres sont estimés par la méthode du maximum de vraisemblance.

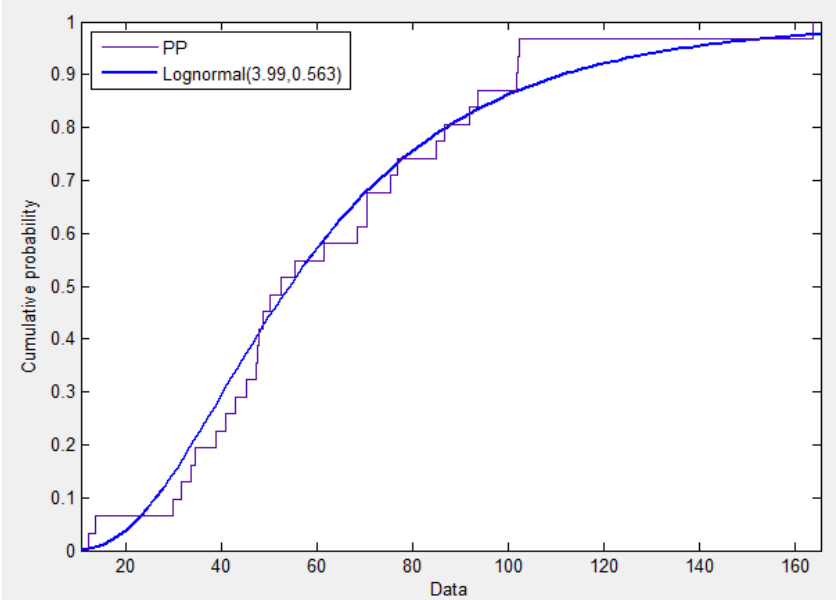


Fig2 : Fonction de répartition empirique des précipitations, et celle de la loi LN

La figure 3 présente les résultats des estimations par bootstrap d'un intervalle de confiance pour le paramètre μ de la loi Log Normal ;nous effectuons 1500 bootstrap.

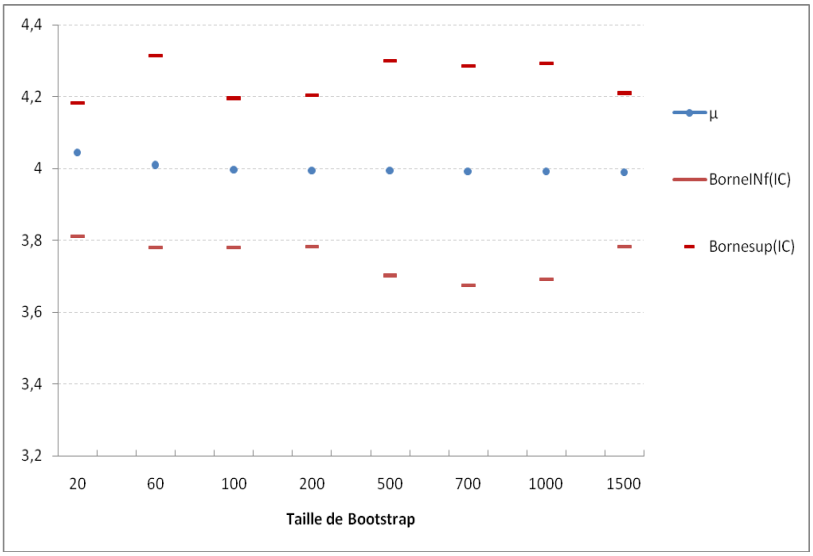


Fig 3 :Estimation de l'IC du paramètre μ par Bootstrap

Du fait de la nature de la corrélation entre les deux variables (corrélation négative), nous sommes amenés à utiliser la copule de Frank. Le résultat d'un test d'ajustement sur l'adéquation de la copule de Frank est représenté dans la figure 4 ; la courbe en rouge correspond à une distribution empirique et celle en bleu correspond à une distribution théorique, ainsi on peut conclure à une bonne adéquation. [2]

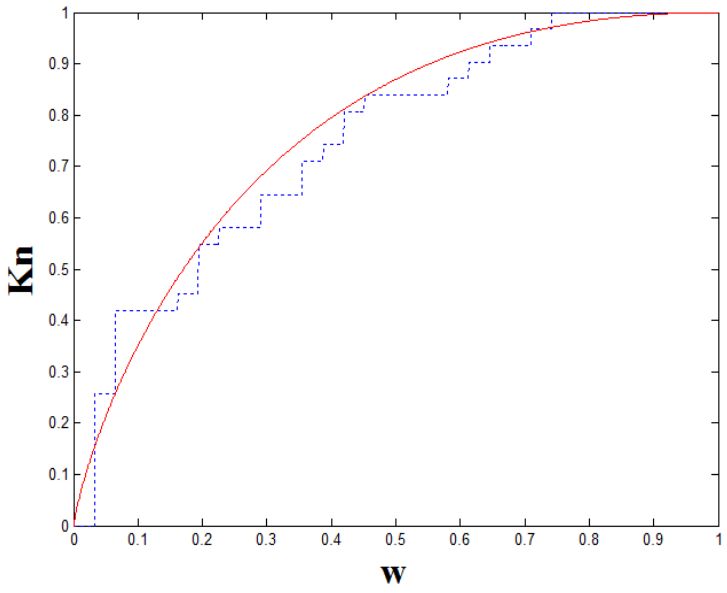


Fig 4 : Ajustement de la copule de Frank

La figure 5 présente les probabilités de dépassement des seuils ; les courbes en bleu correspondent aux quantiles bivariés (iso-probabilité) superposés aux points d'observations. Une température moyenne sur les mois Avril-Mai-Juin de **18.5°C** et un cumul de précipitations de **80mm** durant la même période ont une probabilité de **10 %** d'êtres dépassés.

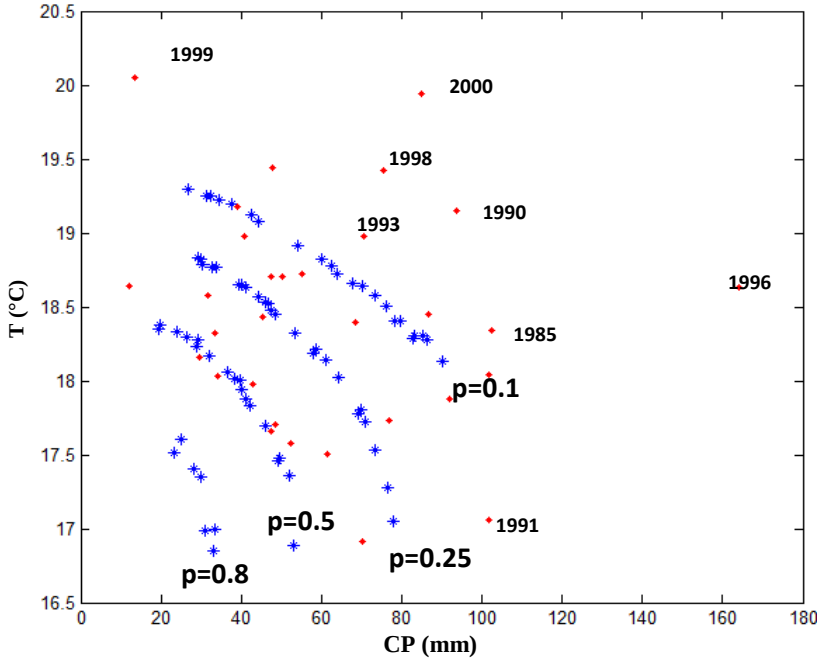


Fig 5 : Quantiles bivariés relatifs à une probabilité de **dépassement**

Pour valider nos résultats, nous désirons effectuer une enquête, qui nous permettra de conclure quant à la pertinence de notre démarche d'une part, et de bien préciser les seuils de déclenchement de l'épidémie d'autre part. Pour ce faire nous adopterons la méthode Q-sort. Nous tenons également à reporter la méthodologie à une échelle mensuelle afin d'améliorer la précision. Enfin, pour mieux apprécier les valeurs du risque nous comptons traduire les résultats en termes de coûts, en introduisant une fonction coût.

Bibliographie:

[1] Andy.S (2002): Science, Precaution, and Practice, Public Health Reports, UK.

[2] Anis.Z et M.Anis.S (2010) : Etude des événements météorologiques extrêmes sur deux sites en Tunisie, PFE, ENIT.

[3] F.Chebana et T.B.M.J. Ouarda(2009) : Multivariate quantiles in hydrological frequency analysis, Environmetrics, INRS-ETE, Québec, Canada.

[4] Genest, C. et Favre, A.C.(2006): Everything you always wanted to know about copula modeling but were afraid to ask, Département de mathématiques et de statistique, Université Laval, Québec, Canada.

[5] Odile. C, Réjean.B, Jacques.L et Wendy.M.S(2006) : Guide d'identification des principales maladies de la vigne, Agriculture et Agroalimentaire, Canada.

Analyse de données d'expression des gènes impliqués dans la polyarthrite rhumatoïde

Sonia Kechaou-Cherif [1-2], Slimane Ben Miled [1], Alia BenKahla [1], Saloua Zaltni-Belhadj [2], Leith Zakraoui [2], Kamel Hamzaoui [3].
1- Groupe de Bio-informatique et de Modélisation Mathéatique à l'Institut Pasteur de Tunis
2 - Service de Rhumatologie, CHU Mongi Slim, La Marsa
3- Département d'immunologie, Faculté de Médecine de Tunis

Résumé:

La polyarthrite rhumatoïde comme décrite par Andreas (2008) est un rhumatisme inflammatoire chronique. C'est une maladie auto-immune d'étiologie incertaine caractérisée par l'inflammation chronique et symétrique des grandes et des petites articulations. Il s'agit du rhumatisme inflammatoire le plus fréquent. Il affecte environ 1% de la population mondiale et touche les deux sexes à l'âge adulte de 40 à 70 ans avec une nette prédominance féminine (sex ratio 1/2.5). Cette maladie a une morbidité importante et un impact socio économique considérable.
Le mécanisme physiopathologique de cette maladie n'est pas encore élucidé et on pense que plusieurs gènes dont l'expression change pourraient être à l'origine de cette maladie.

L'objectif de notre étude est d'identifier ces gènes et de montrer via une méta-analyse leur implication dans la polyarthrite rhumatoïde.

Matériel et Méthodes:
Les données ont été prélevées sur des bases de données publiques à savoir "GEO" et "Array Express". Ces données ont été déjà présentées dans les articles d'Huber (2008) et Koczan (2008). Elles comportaient des données d'expression de gènes prélevés à partir d'échantillon de membrane synoviale de sujets atteints de polyarthrite rhumatoïde et de sujets témoins ainsi que du sang de sujets atteints de polyarthrite rhumatoïde. Nous avons sélectionnées des données brutes uniquement que nous avons normalisées et corrigées par la méthode RMA (Robust Multi-array Average) comme définie par Irizarry (2003). Nous avons procédé par la suite à une analyse de variance en utilisant la méthode ANOVA comme elle a été décrite par Haan (2007) afin de déterminer les gènes différentiellement exprimés. Nous avons également procédé à un classement non supervisé de tous les gènes par la méthode dite "k means clustering" en fonction de l'état du patient (PR ou témoin), du type d'échantillon (sang ou membrane synoviale) et les deux à la fois. Cette dernière méthode a été présentée dans plusieurs articles dont celui de Freyhult (2008).

Résultats:
Nos données comportaient 40 fichiers .CEL relatifs aux 40 sujets inclus. Parmi ces sujets, nous avions 9 témoins et 12 sujets atteints de PR dont l'échantillon a été prélevé à partir de la membrane synoviale et 19 sujets atteints de PR dont l'échantillon a été prélevé à partir du sang.
Chaque fichier .CEL comporte 22283 probe set.
Nous avons pu déterminer les gènes différentiellement exprimés entre les sujets atteints de PR et les sujets témoins: 1375 gènes étaient sur-exprimés chez les sujets PR par rapport aux témoins et 1249 étaient sous-exprimés. Nous avons aussi déterminé les gènes différentiellement exprimés entre les sujets atteints de PR dont l'échantillon a été prélevé à partir de la membrane synoviale et les sujets atteints de PR dont l'échantillon à été prélevé à partir du sang et nous avons trouvé 677 gènes qui étaient sur-exprimés chez les sujets atteint de PR dont l'échantillon provenait de la membrane

synoviale par rapport aux sujets atteint de PR dont l'échantillon provenait du sang et 150 gènes sous exprimés chez ces derniers par rapport aux sujets atteints de PR et dont l'échantillon provenait de la membrane synoviale .
Nous avons également pu diviser les gènes en différents groupes en fonction de leurs niveaux d'expressions en tenant compte de l'état du patient et du type d'échantillon. Le nombre de groupes à été déterminé dans chacun des cas graphiquement à partir d'un plot présentant la Merit Value définie par Yeung (2001) en fonction du nombre de groupes. Nous avons classé nos gènes en 10 groupes en tenant compte de l'état du sujet et en 8 groupes en tenant compte du type d'échantillon. Nous avons, par la suite, étudié le rôle des gènes de chacun de ces groupes en déterminant les pathways dans les quels ils sont impliqués de façon statistiquement significative. Nous avons trouvé que les gènes dont l'expression change entre les états (PR ou témoins) et entre les deux types d'échantillon sont impliqués dans des pathways relatifs à des processus biologiques en rapport avec des phénomènes inflammatoires et des perturbations immunlogiques. En effet, 21 gènes parmi 184 gènes sur-exprimés chez les PR dont l'échantillon provenait de la membrane synoviale (c-à-d 13,9% des gènes présents dans le pathway) ont été retrouvé dans un pathway d'interaction récepteur-membrane extra cellulaire. Ces gènes codent pour le cartilage oligometric matrix protein (COMP), le collagène de type I, III, IV et V, certaines laminines ainsi que certains facteurs de la coagulation tel les thrombospondines et le facteur Willebrand. Ces mêmes gènes ont été retrouvés dans le pathway d'adhésion focale. Douze gènes (c-à-d 8% des gènes présents dans le pathway) ont été retrouvés dans le pathway de la cascade de l'activation du complément et de la coagulation. Par ailleurs, parmi 150 gènes sous-exprimés chez les PR dont l'échantillon provenait de la membrane synoviale 10 gènes étaient impliqués dans le pathway de la régulation de la lignée cellulaire hématopoiétique, 9 dans le pathway de la cytotoxicité médiée par les cellules natural killer et 10 dans le pathway des voies de signalisation des interleukines. D'autre part, nous avons montré que les gènes dont l'expression ne change pas d'une condition à une autre sont imliqués dans des pathways relatifs à des processus biologiques physiologiques.

Mots clé: analyse de données de micro-array, polyarthrite rhumatoïde.

Abstract:

Rheumatoid arthritis (RA), as described by Andreas (2008), is a chronic inflammatory and systemic autoimmune disease characterized by a chronic inflammation of synovial joints which can cause severe morbidity and disability. It represents the most frequent inflammatory rheumatism. It affects women more than men (sex ratio=1/2,5) and about 1% of the world's population.
The pathogenesis of RA is complex and not fully understood until now. It was shown that the expression of several genes change in RA patients versus control subjects. These genes likely have a role in the pathogenesis mechanism of the RA.

The aim of our study is to identify these genes and to show through a meta-analysis evidences supporting their involvement in RA.

Materials and methods :
Our data were extracted from the public databases "GEO" and "Array Express". Selected data were presented by Huber and co-authors (2008) and Koczan and co-authors (2008). They included gene expression data provided from analyses of synovial membrane of subjects suffering from RA and control ones and from blood of RA patients. We have normalized and corrected raw data using RMA (Robust Multi-array Average) as defined by Irizarry (2003). Then, we have conducted an analysis of variance using the ANOVA method as it was described by Haan (2007) to identify

differentially expressed genes. We also conducted an unsupervised classification of all genes using "k means clustering" and took into account each time patient status (RA or control), sample type (blood or synovial membrane) and both at the same time. The latter method has been presented in several papers including the Freyhult (2008).

Results :
Our data included 40 ".CEL" files related to the 40 subjects. Among these subjects, we had 9 control, 12 RA patients whose sample was taken from the synovial membrane and 19 RA patients whose sample was taken from blood. Each ".CEL" file include 22283 probes.
We have identified genes differentially expressed between RA patients and control subjects: 1375 genes were over-expressed in RA patients and 1249 were under-expressed. We also determined the genes differentially expressed between synovial membrane samples and blood samples : 677 genes were over-expressed in synovial membrane samples of RA patients and 150 genes were under-expressed in blood samples of RA patients.
Moreover, we have divided the genes into different groups according to their expression levels and taking into account the patient status and sample type. The groups number was determined graphically based on a plot of the Merit Value versus the number of groups. We classify our genes into 10 and 8 groups taking into account, respectively, the subject status and sample type. Finally, we studied the role of genes in each of these groups by determining the pathways in which they are significantly involved. Genes differentially expressed between the different conditions are involved into pathways relating inflammatory and immunological processes which can likely have a role in the pathogenesis mechanism of the RA. However, genes not differentially expressed are essentially involved into pathways representing physiological phenomena.

Key words: microarray data analysis, rheumatoid arthritis.

Bibliographie:

[1]Andreas, K. Lübke, C. Häupl, T. et al (2008) Key regulatory molecules of cartilage destruction in rheumatoid arthritis: an in vitro study. *Arthritis Research & Therapy,* 10, 1, R9.

[2]Huber, R. Hummert, C. Gausmann, U. et al (2008) Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. *Arthritis Reseach & Therapy,* 10, 4, R98.

[3]Koczan, D. Drynda, S. Hecker, M. et al (2008) Molecular discrimination of responders and nonresponders to anti-TNF alpha therapy in rheumatoid arthritis by etanercept. *Arthritis Research & Therapy,* 10, R50.

[4]Irizarry, RA. Hobbs, B. Collin, F. et al (2003) Exploration, noramalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 2, 249-264.

[5]Haan, JR. Wehrens, R. Bauerschmidt, S. et al (2007) Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics,* 23, 2, 184–190.

[6]Freyhult, E. Landfors, M. Önskog, J. et al (2010) Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *Bioinformatics*, 11, 503.

[7]Yeung, KY. Haynor, DR. Ruzzo, WL. (2001) Validating Clustering for gene expression data. *Bioinformatics,* 17, 4, 309-318.

Lecture Probabiliste du Cycle Boursier Tunisien : Proposition d'un modèle à trois états avec changements de régimes markoviens et dépendance à la durée

Dr. Adel KARAA
Professeur à l'Institut Supérieur de Gestion de Tunis
e-mail : adel.karaa@isg.rnu.tn
adelkaraa@yahoo.fr
Tel : 00 216 98 542 585

Emna MAHAT
Doctorante
e-mail : *emna_mahat@hotmail.fr*
Tel : 00 216 22 225 216

Azza BEJAOUI
Doctorante
e-mail : bjaouiazza2@yahoo.fr
bjaouiazza2@hotmail.fr
Tel : 00 216 25 935 675

Introduction

Depuis plusieurs années, la recherche de la nature des dynamiques boursières a suscité l'intérêt de plusieurs chercheurs (e.g., Fama, 1965 ; Mandelbrot, 1963). L'attention accordée à ce sujet a abouti à un consensus selon lequel les séries chronologiques partagent un éventail de propriétés statistiques appelées les faits stylisés (*stylized facts*). Une bonne compréhension de ces caractéristiques demeure un enjeu scientifique majeur pour les économètres et financiers afin de mieux appréhender le comportement des rendements boursiers. Parmi les faits stylisés découverts dans les séries des rendements à haute fréquence, on note la distribution de rendements à queues épaisses '*fat tails*' (e.g., Engle, 1982 ; Clark, 1973 ; Mandelbrot, 1963), le regroupement de volatilité '*volatility clustering*' (e.g., Mandelbrot, 1963), la mémoire longue '*long memory*' (e.g., Granger, 2001).

Par la même occasion, l'examen des séries de rendements boursiers a révélé la présence d'une forte composante cyclique qui indique que le marché financier passe par une tendance suffisamment longue de hausse ou baisse des cours. Cette succession de phases ascendantes et descendantes caractérisant, le plus souvent, l'évolution des cours boursiers est connu par les marchés haussiers (*Bull Markets*) et baissiers (*Bear Markets*). L'identification de ces phénomènes reste, selon Chen et Shen (2007), une préoccupation majeure pour non seulement les intervenants sur les marchés boursiers mais aussi les universitaires. Malgré cet accord étendu sur cette importance, il n'y a toujours pas de définition largement acceptée dans la littérature. Traditionnellement, Sperandeo (1990) prétendait qu' « *un marché haussier est un mouvement du prix à la hausse à long terme caractérisé par une série des hauts intermédiaires plus élevés interrompue par une série des bas intermédiaires plus élevés* » alors qu'un marché baissier « *est une tendance à la baisse à long terme caractérisée par des bas intermédiaires plus faibles interrompue par des hauts intermédiaires plus faibles* ». Pour leur part, Chauvet et Potter (2000) pensent, en se référant à la terminologie du marché boursier, qu'un marché baissier (haussier) correspond à des périodes où les prix du marché sont généralement décroissants (croissants). Pagan et Sossounov (2003) déclarent qu'une augmentation (diminution) des cours boursiers supérieure (inférieure) à 20% ou 25% peut qualifier respectivement le marché de haussier ou baissier. D'une manière plus générale, Chen et Shen (2007) avancent que les marchés haussiers (baissiers) sont définis en termes de rendements dépassant (ne dépassant pas) un certain seuil dans un mois donné.

Malgré ce foisonnement de définitions cherchant à éclaircir les termes 'marchés haussiers' et 'marchés baissiers', la bonne compréhension de ces concepts reste discutable. Néanmoins, comme le rappellent Maheu et *al*. (2010), il est important de les extraire à partir des données en vu d'analyser leurs propriétés et de s'en servir éventuellement dans la prise de décisions d'investissement. Pour cela, deux différentes méthodes permettent la détection des marchés haussiers et baissiers à savoir : les algorithmes de datation[1] (*Dating Algorithms*) et les modèles à Changement de Régimes (*Regime-*

[1]Les algorithmes de datation repose sur une évaluation *ex post* des pics et des creux de l'indice boursier en vu de définir les marchés haussiers (d'un creux à un pic) et baissiers (d'un pic à un creux).

Switching Models). Bien que la première méthode ait très souvent servie les chercheurs (e.g., Gonzalez et *al*., 2006 ; Pagan et Sossounov, 2003 ; Biscarri et de Garcia, 2002) à délimiter concrètement les périodes de hausse et baisse du marché boursier, elle ne pourrait pas néanmoins être utilisée pour l'inférence statistique et la prévision exigeant plus d'informations à partir de la distribution des rendements (Maheu et *al*., 2009). Pour pallier à ces limites, les modèles à changement de régimes markoviens, introduits pour la première fois par Goldfeld et Quandt (1973) et développés par Hamilton (1989), ont réussi énormément à reproduire statistiquement les dynamiques boursières. L'idée sous-jacente à ce type de modèles est que le marché boursier passe d'un régime (marché haussier) à un autre (marché baissier) suivant une variable d'état latente représentée par une chaîne de Markov notée S_t. Les résultats dégagés de l'application des modèles à changement de régimes markoviens aux rendements boursiers indiquent que le marché haussier se caractérise par des rendements élevés et une volatilité faible tandis que le marché baissier montre des rendements faibles et une volatilité élevée (e.g., Chen 2007 ; Tu, 2006).

Maheu et McCurdy (2000) introduisent la notion de la dépendance à la durée[2] dans le modèle à changement de régimes markoviens (*Duration-Dependent Markov Switching Model*) afin de capter la structure non-linéaire dans la moyenne conditionnelle et la variance conditionnelle dans les rendements boursiers américains. La prise en compte de cette notion permet de détecter certains faits tels que le *clustering* de volatilité, le retour à la moyenne et les caractéristiques cycliques non linéaires dans les rendements. Les résultats de cette étude révèlent, en plus des caractéristiques déjà citées, une dépendance à la durée négative sur les marchés haussiers et baissiers. D'une manière plus explicite, la probabilité de changer d'un état à un autre décroit avec la durée d'être dans cet état. Par ailleurs, Maheu et McCurdy (2000) affirment que les meilleurs gains du marché semblent avoir lieu au début d'un marché haussier alors que la volatilité croît avec la durée durant un marché baissier. Plus récemment et en reprenant le modèle de Pelagatti (2001), Chen et Shen (2007) concluent que le Japon, le Hong Kong et la Corée du Sud ont certaines caractéristiques communes à savoir : une dépendance à la durée (absence de dépendance) dans le marché baissier (haussier). En revanche, pour le Singapore et le Taïwan, il existe une dépendance à la durée dans les deux marchés haussiers et baissiers.

Dans cet article, nous cherchons à étudier les dynamiques boursières régissant le marché financier tunisien tout en proposant un modèle à trois états avec changements de régimes markoviens appliqué à la série de rendements hebdomadaires de l'indice boursier TUNINDEX durant la période allant du 07/01/1998 au 19/08/2010.

1. Modèle

Nous nous plaçons ici dans le cadre général des modèles à changements de régimes markoviens. Pour les besoins de notre modèle, nous supposons qu'à chaque instant t, la série des rendements peut appartenir à un régime donné parmi trois régimes possibles[3]. Ces derniers étant eux-mêmes non observables, nous les exprimons par l'intermédiaire d'une variable latente S_t^* qui prend les valeurs {1, 2, 3}. La dynamique de transition entre les trois régimes est décrite par un processus semi markovien homogène. Ce type de processus présente l'avantage d'autoriser, contrairement au processus markovien utilisé généralement dans les modèles à changement de régimes (Hamilton, 1989), une dépendance des intensités de transition à l'égard de l'ancienneté des régimes qui est exprimée par une deuxième variable latente $D_{S_t^*}$ désignant, la durée, ou encore le nombre de périodes successives récemment passés dans un même régime :

$$D_{S_t^*} = \begin{cases} D_{S_{t-1}^*} + 1 & \text{si} \quad S_t^* = S_{t-1}^* \\ 1 & \text{sinon} \end{cases}. \qquad (1.1)$$

[2]Le modèle initié par Hamilton (1989) ne prend pas en considération la notion de la dépendance à la durée dans les états. Pour cela, Durland et McCurdy (1994) modifient la structure des probabilités de transition pour les permettre d'être dépendant de la durée (*Duration-Dependent Transition probabilities*).

[3] L'hypothèse de trois régimes est justifiée *a posteriori*.

Pour chaque régime i (i = 1, 2, 3), les probabilités de transitions sont spécifiées conditionnellement à $D_{S^*_{t-1}}$ de la manière suivante :

$$p_{ij}^{d} = Pr\left(S_t^* = j / S_{t-1}^* = i ; D_{S^*_{t-1}} = d\right) = \frac{\exp\left[\lambda_1^{ij} + \lambda_2^{ij}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right]}{1+\exp\left[\lambda_1^{ij} + \lambda_2^{ij}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right] + \exp\left[\lambda_1^{ik} + \lambda_2^{ik}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right]} \tag{1.2}$$

$$p_{ik}^{d} = Pr\left(S_t^* = k / S_{t-1}^* = i ; D_{S^*_{t-1}} = d\right) = \frac{\exp\left[\lambda_1^{ik} + \lambda_2^{ik}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right]}{1+\exp\left[\lambda_1^{ij} + \lambda_2^{ij}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right] + \exp\left[\lambda_1^{ik} + \lambda_2^{ik}\left(d\, I_{(d\le\tau)} + \tau\, I_{(d>\tau)}\right)\right]}^{4} \tag{1.3}$$

Dans le cadre général des modèles de durées, les probabilités p_{ij}^{d} et p_{ik}^{d} peuvent désigner des fonctions de hasard et traduisent respectivement les probabilités instantanées de quitter le régime *i* au profit des régimes *j* et *k* sachant qu'on ait jusque là dans le même régime *i* depuis *d* périodes. Etant donné que l'on associe à chaque changement de régime deux destinations possibles (concurrentes et exclusives), le modèle envisagé peut être considéré comme étant un modèle de durées à risques concurrents. Néanmoins, il y a lieu de rappeler que durant d'autres périodes, il se peut qu'il n'y ait pas un changement de régime. Si tel est le cas, on parle ainsi d'une situation de persistance dans les régimes qui est considérée comme étant une troisième destination possible, concurrente aux deux premières déjà citées, et qui est désignée formellement par la probabilité suivante :

$$p_{ii}^{d} = 1 - p_{ij}^{d} - p_{ik}^{d} \tag{1.4}$$

Au niveau de chaque régime i, on parle d'une dépendance des fonctions de hasard (ou encore les probabilités de transitions) vis-à-vis l'ancienneté du régime lorsqu'au moins un des coefficients λ_2^{ij} et λ_2^{ik} est significativement différent de zéro. Cette dépendance est dite positive (respectivement, négative) si de plus ces coefficients sont positifs (respectivement, négatifs). Dans ces conditions, les fonctions de hasard se montrent comme croissantes (respectivement, décroissantes) en fonction de l'ancienneté du régime *i*, auquel elles associent par la même occasion un niveau de persistance qui est décroissant (respectivement, croissant) au fil du temps. Par hypothèse, nous supposons que ce niveau de persistance devient constant au-delà d'un horizon de τ périodes. Le paramètre τ désigne, comme l'indique Durland et McCurdy (1994) et Maheu et McCurdy (2000), la mémoire du processus semi markovien[5].

D'une manière plus formelle, le modèle proposé que nous indiquions par DD(τ)-MS(3)-AR(L)-GARCH-M(1,1)[6] se présente comme suit :

$$R_t = \alpha^{\left(S_t^*\right)} + \delta^{\left(S_t^*\right)} \ln\left(d_{S_t^*}\right) + \sum_{l=1}^{L} \beta_l^{\left(S_{t-l}^*\right)}\left[R_{t-l} - \alpha^{\left(S_{t-l}^*\right)} - \delta^{\left(S_{t-l}^*\right)} \ln\left(d_{S_{t-1}^*}\right)\right] + \eta^{\left(S_t^*\right)} h_t(S_t) + \varepsilon_t \quad , \tag{1.5}$$

---

[4] $I_{(d\le\tau)}$ et $I_{(d>\tau)}$ désignent deux variables indicatrices qui se présentent respectivement comme suit :

$$I_{(d\le\tau)} = \begin{cases} 1 & \text{si } d \le \tau \\ 0 & \text{sinon} \end{cases} \quad et \quad I_{(d>\tau)} = \begin{cases} 1 & \text{si } d > \tau \\ 0 & \text{sinon} \end{cases}$$

[5] Etant donné que ce paramètre ne peut prendre que des valeurs discrètes, il est déterminé par une méthode de balayage qui consiste à maximiser la fonction de vraisemblance, fonction que nous retenons pour estimer tous les paramètres du modèle.

[6] C'est une démarche statistique qui vient justifier la pertinence de la spécification retenue. Elle consiste à étudier dans une première étape la stationnarité de la série des rendements hebdomadaires de TUNINDEX moyennant les tests de Duckey-Fuller augmenté et de Perron. De déterminer dans une deuxième étape l'ordre des processus AR retenu pour décrire la dynamique des rendements moyennant les critères d'information AIC et BIC. Cette procédure est complétée par l'étude des fonctions d'auto-corrélation simple et partielle et des résultats du test Ljung-Box portant sur les données filtrées (i.e., la série des résidus issus de l'ajustement de la série des rendements par un modèle ARMA). Le test du Multiplicateur de Lagrange (LM) est utilisé dans une dernière étape pour tester la présence des effets ARCH dans la série des rendements filtrés. Le choix de trois régimes est justifié a posteriori en faisant appel au test de rapport de vraisemblance adapté par Hansen (1992, 1996) et Garcia (1998) aux modèles à changements de régimes markovien.

$$h_t(S_t) = \gamma_0^{(S_t^*)} e^{\gamma_1^{(S_t^*)} \ln\left(d_{S_t^*}\right)} + \gamma_2^{(S_{t-1}^*)} (\tilde{\varepsilon}_{t-1}^{(S_{t-1}^*)})^2 + \gamma_3 \tilde{h}_{t-1} \quad , \tag{1.6}$$

$$\tilde{\varepsilon}_{t-1}^{(S_{t-1}^*)} = R_{t-1} - E_{t-2}\left[R_{t-1}/S_{t-1}^*\right], \tag{1.7}$$

$$E_{t-2}\left[R_{t-1}/S_{t-1}^*\right] = \frac{\sum_{n=1}^{N} E_{t-2}\left[R_{t-1}/S_{t-1}=n, Y_{t-2}\right] \Pr\left(S_{t-1}=n/Y_{t-1}\right) I_{(S_{t-1}^*)}}{\sum_{n=1}^{N} \Pr\left(S_{t-1}=n/Y_{t-1}\right) I_{(S_{t-1}^*)}} \quad , \tag{1.8}$$

$$\tilde{h}_{t-1}^2 = \sum_{n=1}^{N} h_{t-1}^2(n) \Pr\left(S_{t-1}=n/Y_{t-1}\right) \quad , \qquad \varepsilon_t \to NID(0, h(S_t)), \tag{1.9}$$

Avec :

- R_t : Le rendement de l'indice du marché à l'instant t ;
- Y_t : $(R_t, R_{t-1}, \ldots, R_1)$, le vecteur des rendements de l'indice du marché pour les périodes t, t-1, … , 1 ;
- S_t :{1, 2, 3, …, N}, une variable latente à '*N états*' suivant une chaîne de Markov d'ordre 1 qui se caractérise par la propriété suivante **:**

$$Pr(S_t = n/S_{t-1} = m, \ldots, Y_{t-1}) = Pr(S_t = n/S_{t-1} = m, Y_{t-1}) = p_{nm} \quad \text{avec n, m = \{1, 2, 3, …, N\}} \tag{1.10}$$

$\Pr\left(S_t = n/Y_{t-1}\right)$; n = {1, 2, …, N}, vecteur associé des probabilités filtrées exprimant pour tout instant t, les probabilités conditionnelles d'occurrence des 'états', sachant toute l'information jusqu'au temps t-1.

Dans notre modèle, chaque réalisation de la variable S_t fait référence à une trajectoire possible de longueur L+τ du processus DD(τ)-MS(3)-AR(L)-GARCH-M(1,1). Au plan de la construction, on identifie la trajectoire à un vecteur de situations mensuelles retraçant la dynamique des variables S_t^*, et $D_{S_t^*}$ sur des épisodes de temps de L+τ périodes (semaine)[7]. La variable $I_{(S_t^*)}$ est retenue pour indiquer les trajectoires s'achevant par une même situation S_t^* en fin d'épisode.

Dans l'équation (1.5) l'évolution du rendement de l'indice du marché est supposée être régie par un processus autorégressif d'ordre L. Dans notre cas, c'est le test de Ljung et Box (1978) appliqué aux résidus standardisés de cette équation qui est privilégié pour se prononcer sur l'ordre autorégressif L à retenir. Il est à noter que les termes d'erreur de l'équation de rendement ne peuvent être observés étant donné que les régimes sont eux même non observables. Nous suggérons dès lors, à l'instar de Dueker (1997) et Maheu & Mc Curdy (2000), de retenir pour ce test les résidus espérés standardisés[8] :

$$\sum_{J=1}^{N} \frac{R_t - E\left[R_t/S_t = n, Y_{t-1}\right]}{\sqrt{h_t(S_t)}} \Pr\left(S_t = n/Y_{t-1}\right) \tag{1.11}$$

D'ailleurs, c'est pour cette même raison que nous avons exprimé les termes autorégressifs de l'équation de volatilité en fonction des termes $\tilde{\varepsilon}_{t-k}$ et $\tilde{h}_{t-1}^{(S_{t-1}^*)}$ plutôt que ε_{t-k} et $h_{t-1}(S_{t-1})$, respectivement.

Les équations (1.5) et (1.6) sont exprimées en fonction de la variable $D_{S_t^*}$ afin de tenir compte de la dépendance respective des moyennes et variances conditionnelles des rendements vis-vis de l'ancienneté des régimes. La forme exponentielle retenue dans l'équation (1.6) est de nature à assurer la positivité de la variance conditionnelle. In fine, le recours à des relations explicites entre la moyenne conditionnelle et la variance conditionnelle au niveau de l'équation (1.5) permet d'élucider la nature de l'arbitrage entre le rendement et le risque (*trade-off* rendement-risque) au niveau de chaque régime.

2. Résultats et Interprétation

Conviés à chercher, sur le plan empirique, des spécifications adéquates pour reproduire la dynamique non linéaire de l'indice TUNINDEX, pour la période d'étude, nous avons testé trois modèles qui se distinguent principalement par le nombre de régimes. Pour cela, nous avons estimé un modèle (M1) à un seul régime avec une spécification de type AR(2) pour l'équation de rentabilité et

[7] Cela suppose que la structure à trois régimes que nous avons retenue par hypothèse dans le cadre de notre modèle permet de décrire entièrement et sans ambiguïté la chronique des rendements mensuels de l'indice TUNINDEX.

[8] Dans ces conditions, les résultats du test de Ljung et Box (1978) ne peuvent être utilisés qu'à titre indicatif étant donné que la distribution asymptotique de la statistique est inconnue.

une spécification du type GARCH(1,1) pour l'équation de volatilité. Le caractère significatif de tous les coefficients estimés (tableau 1) nous conduit à rejeter l'hypothèse de marche aléatoire pour décrire l'évolution des fluctuations des rendements hebdomadaires de l'indice boursier TUNINDEX. Le niveau significatif du coefficient β_1 (relatif à la partie autorégressive du processus GARCH) nous donne déjà la conviction du caractère non linéaire de ce processus. S'agissant d'une paramétrisation quadratique de la variance conditionnelle, les spéciations de type GARCH se limitent au traitement de la non linéarité au niveau de l'équation de la volatilité sans se donner les moyens de l'étudier au niveau de l'équation de rentabilité. Dans cette optique, deux autres modèles sont retenus pour pousser plus loin l'examen des effets de la non-linéarité conjointement sur la rentabilité et la volatilité de l'indice à savoir : un modèle (M2) avec deux régimes et un modèle (M3) avec trois régimes. Pour ces deux modèles, nous avons tenu compte de la dépendance des intensités de transitions à l'égard de l'ancienneté des régimes. Au-delà de l'étude de la non-linéarité, ces deux derniers mettent en place les bases d'une lecture probabiliste du cycle boursier tunisien. A ce niveau de l'analyse, ceci revient à mettre en avant le caractère cyclique des fluctuations de l'indice TUNINDEX tout en traçant des trajectoires selon lesquelles la chronique des rendements soit décrite par une succession de phases ascendantes, descendantes, ou encore sans tendance dans le cas des modèles à trois régimes.

La lecture de la troisième colonne relative au modèle (M2)[9] permet de distinguer entre deux états du marché qui s'opposent fortement du point de vue du niveau, de la volatilité et de la persistance des rendements. Les signes opposés que nous enregistrons au niveau des termes constants de l'équation de la rentabilité font déjà état de l'existence de deux régimes qui font allusion à une rentabilité élevée et faible, respectivement. Par la même occasion, les graphiques présentés ci-dessous (Fig.1) ne font qu'approuver la présence d'une différence substantielle entre les deux régimes en termes non seulement de variance conditionnelle, rentabilité conditionnelle mais aussi de persistance. En effet, on remarque que le régime 2 se présente comme un état à volatilité faible, rendement faible qui croît avec la durée et tend à être finalement persistant avec la durée. En revanche, le régime 1 montre une volatilité élevée, une rentabilité élevée ayant tendance à décliner à travers le temps ainsi qu'une persistance qui s'affaiblit avec la durée.

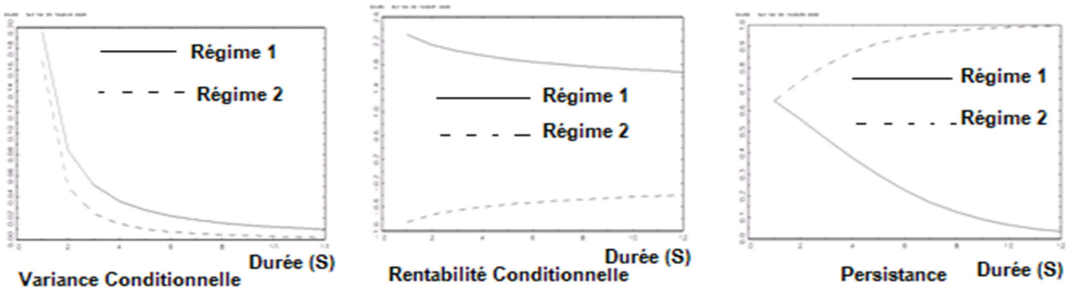


Fig. 1 : Modèle (M2), Variance Conditionnelle, Rentabilité Conditionnelle & Persistance

Il est claire que toutes ces constations concourent, en conformité avec la littérature (e.g., Henry, 2009 ; Maheu et McCurdy, 2000), à retenir deux régimes différents pour décrire la chronique des rendements de l'indice. Néanmoins, le régime 1 que nous qualifions de 'haussier' (*Bull State*) se caractérise par des niveaux de rentabilité et volatilité élevés mais une persistance relativement plus faible que ceux observés dans le deuxième régime. Ce dernier nous le considérons comme étant un régime 'baissier' (*Bear State*) vu qu'il se caractérise par un niveau de rentabilité faible.

[9]Le résultat du test du rapport de vraisemblance est ici sans ambiguïté en faveur du modèle à deux régimes.

Tableau 1 : Résultats d'Estimation des Modèles (M1), (M2) et (M3) par la Méthode du Maximum de Vraisemblance

Paramètres		Modèle (M1) : AR(2)-GARCH-M (1,1)	Modèle (M2) : DD(τ)-MS(2)-AR(2)-GARCH-M (1,1)	Modèle (M3) : DD(τ)-MS(3)-AR(2)-GARCH-M (1,1)
Equation de rentabilité	$\alpha^{(1)}$	0.0018 (0.0032) ***	2.3054 (0.0586) ***	0.9323 (0.0201) ***
	$\alpha^{(2)}$		-0.8562 (0.0443) ***	-0.6733 (0.0205) ***
	$\alpha^{(3)}$			-2.1525 (0.0315) ***
	$\beta_1^{(1)}$	-0.0160 (0.0017) ***	0.0557 (0.0145) ***	-0.1627 (0.0119) ***
	$\beta_1^{(2)}$		0.0620 (0.0151) ***	0.2574 (0.0088) ***
	$\beta_1^{(3)}$			0.0077 (0.0149)
	$\beta_2^{(1)}$)	0.0422 (0.0128) **	0.3549 (0.0064) ***	0.6183 (0.0065) ***
	$\beta_2^{(2)}$)		0.1464 (0.0086) ***	0.3689 (0.0057) ***
	$\beta_2^{(3)}$			0.6382 (0.0133) **
	$\delta^{(1)}$		-0.2545 (0.0262) ***	0.3349 (0.0146) ***
	$\delta^{(2)}$		0.1840 (0.0241) ***	0.5523 (0.0135) ***
	$\delta^{(3)}$			-0.2095 (0.0679) ***
	$\eta^{(1)}$		0.4811 (0.0171) ***	0.1670 (0.0508) ***
	$\eta^{(2)}$		-0.2416 (0.0132) ***	-0.1028 (0.0653) *
	$\eta^{(3)}$			-0.1055 (0.0262) ***
Equation de volatilité	$\gamma_0^{(1)}$	0.0002 (6.6E -06) **	0.1959 (0.0110) ***	0.0069 (0.0004) ***
	$\gamma_0^{(2)}$		0.1681 (0.0119) ***	0.0172 (0.0015) ***
	$\gamma_0^{(3)}$			0.9280 (0.1381) ***
	$\gamma_1^{(1)}$		-1.2161 (0.0612) ***	-0.5525 (0.1419) ***
	$\gamma_1^{(2)}$		-1.7551 (0.0509) ***	-1.5026 (0.1658) ***
	$\gamma_1^{(3)}$			-2.4196 (0.1665) ***
	$\gamma_2^{(1)}$	0.2926 (0.0213) **	0.5722 (0.0370) ***	0.0327 (0.0037) ***
	$\gamma_2^{(2)}$		0.2748 (0.0146) ***	0.0034 (0.0012) ***
	$\gamma_2^{(3)}$			0.8364 (0.0839) ***
	γ_3			0.0379 (0.0052) ***
			0.0530 (0.0060) ***	
Probabilités de transition	$\lambda_1^{(1,2)}$		-0.9626 (0.2987) ***	-1.7706 (1.4095)
	$\lambda_1^{(1,3)}$			-0.6995 (1.9530)
	$\lambda_1^{(2,1)}$			
	$\lambda_1^{(2,3)}$		-0.1513 (0.1550)	-1.5090 (0.7161) **
				-4.0223 (1.1103) ***
	$\lambda_1^{(3,1)}$			1.4698 (0.4050) ***
	$\lambda_1^{(3,2)}$			1.4713 (0.4418) ***
	$\lambda_2^{(1,2)}$		0.3646 (0.1065) ***	1.6564 (1.0275) *
	$\lambda_2^{(1,3)}$			-1.0387 (1.7337)
	$\lambda_2^{(2,1)}$			
	$\lambda_2^{(2,3)}$		-0.4455 (0.0929) ***	1.9766 (0.4227) ***
				-3.3116 (0.9082) ***
	$\lambda_2^{(3,1)}$			0.0633 (0.1571)
	$\lambda_2^{(3,2)}$			-0.1113 (0.1987)
Log-vraisemblance			-520.021458	-139.2612266
Nombre d'observations			666	666
QM(5)				3.315 (n.s)
QM(10)				8.736 (n.s)
QM(15)				12.197 (n.s)
QM(20)				19.532 (n.s)
QM(30)				28.523 (n.s)

Notes : - (*n.s*) : non significatif, -* : Niveau de signification de 10%, - ** : Niveau de signification de 5%, - *** : Niveau de signification de 1%. - Les valeurs entre parenthèses représentent les écarts-types.

Le modèle (M3) est retenu pour étendre le champ de découpage de la chronique des rendements dans une optique à trois régimes. Les résultats de l'estimation de ce modèle par la méthode du maximum de vraisemblance sont rapportés dans la quatrième colonne du tableau 1. Une première lecture des différents coefficients semble indiquer à première vue trois régimes distincts[10]. D'après la figure 2, nous pouvons constater l'émergence de deux régimes diamétralement opposés associant les caractéristiques suivantes : rentabilité positive (resp. négative), croissante (resp. décroissante) en fonction de l'ancienneté du régime et une volatilité très faible (resp. élevée) pour le régime 1 (resp. régime 3). Ces caractéristiques nous donnent le droit de qualifier, conformément à la définition avancée par Chauvet et Potter (2000), le régime 1 (resp. régime 2) de haussier (resp. baissier)[11]. L'autre régime que nous dénotons par « régime 2 » semble se présenter, d'après la figure 2, comme un régime de centre avec des niveaux intermédiaires de rentabilité et de volatilité. Les rendements initialement négatifs au début de période et qui ont tendance à croitre d'une manière progressive avec le temps pour atteindre des niveaux de rentabilité presque nuls ne peuvent que conforter l'idée soutenue par Guidolin et Timmermann (2005) et qui consiste à considérer ce régime comme régime « normal » décrivant un retour à la moyenne.

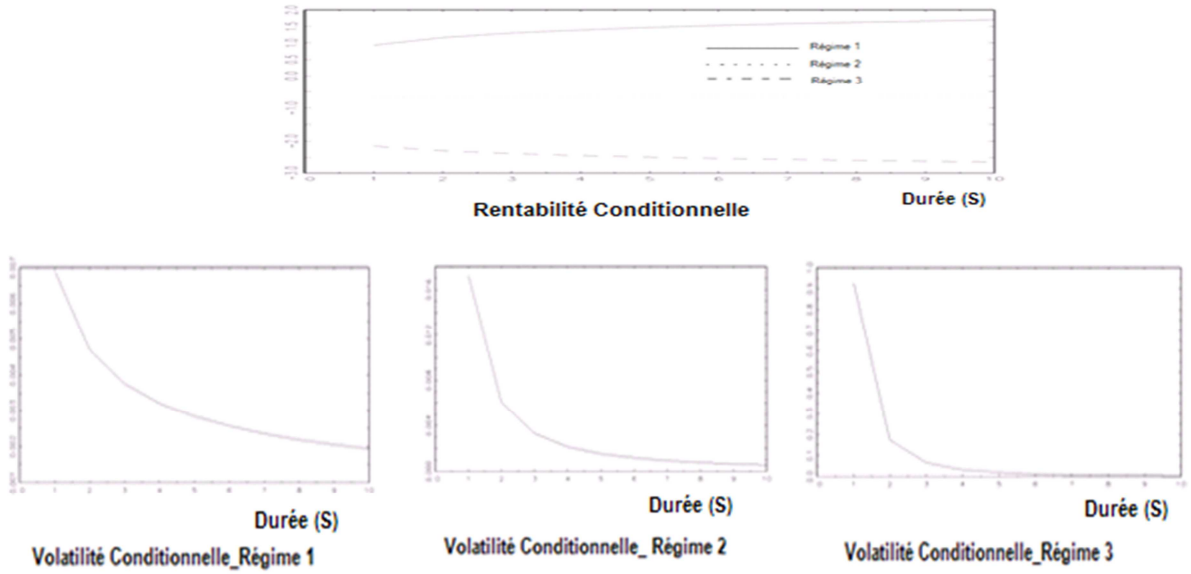


Fig. 2 : Modèle (M3), Variance Conditionnelle, Rentabilité Conditionnelle

Au même titre que le modèle précédent, tous les régimes semblent montrer une dépendance négative de la volatilité à l'égard de la durée. Ceci n'est pas le cas lorsque cette dépendance est abordée au niveau des probabilités de transition. En effet, les différences enregistrées au niveau du signe, de l'amplitude et du niveau de signification des différents paramètres estimés, $\lambda_2^{(i,j)}$ et $\lambda_2^{(i,k)}$, ne peuvent que conduire vers des schémas d'interprétation différents selon les différents cas d'espèce. Pour nous simplifier la tâche, nous nous sommes reportés à la figure 3 qui retrace la dynamique de transition entre les régimes en fonction de la durée.

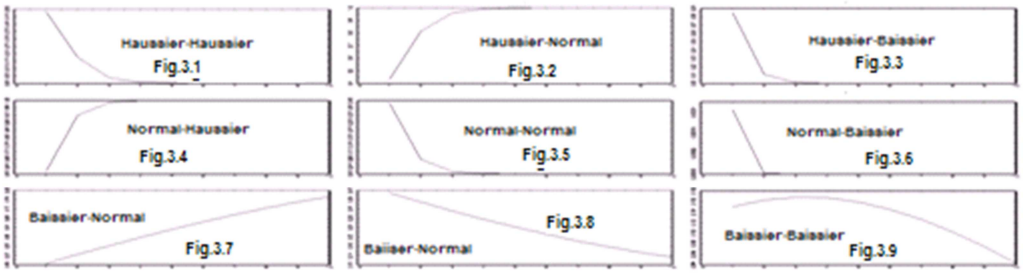


Fig. 3 ***: Evolution des probabilités de persistance et de transition selon la durée pour le Modèle*** **(M3)**

[10]D'ailleurs, le résultat du test du rapport de vraisemblance est ici sans ambiguïté en faveur du modèle à trois régimes.

Les trois tracés retenues sur la diagonale (Fig. 3.1 ; 3.5 ; 3.9) se réfèrent à des situations de persistance des régimes 1, 2 et 3, respectivement. En revanche, les six autres illustrent des situations de migration indiquant le passage d'un régime à un autre. A première vue, nous pouvons remarquer que l'allure de chacune des courbes est soit croissante ou décroissante en fonction de l'ancienneté du régime (exprimée en semaines). Ce premier résultat montre a *priori* l'existence d'une dépendance à la durée qui est de nature à remettre en cause le caractère constant des probabilités de transition.

D'après les figures 3.1 et 3.2, au début de chaque épisode, le régime haussier semble avoir uniquement la chance de persister ou de céder la place au régime normal avec des probabilités respectives de l'ordre de 0.5 et 0.4. Au fur et à mesure que l'épisode en question avance, cette deuxième alternative commence à devenir la seule alternative possible avec des probabilités avoisinant la valeur de 1 à partir de la deuxième semaine. Ce même schéma peut être partiellement retenu pour le régime normal qui semble lui aussi avoir à peu prés 40% de chance de persister et 60% de migrer vers l'état haussier. Cependant, et contrairement au cas précédent, c'est plutôt la probabilité de persistance qui a tendance à décroitre au profit de la probabilité de transition vers le régime haussier. Là encore, cette dernière commence à avoisiner la valeur de 1 au bout de 2 semaines. La situation est tout autre pour le régime baissier. En effet, les chances de persistance sont ici très faibles bien qu'elles ont tendance à augmenter, pour atteindre un niveau maximum de l'ordre de 11% durant les trois premières semaines. Le risque de migration semble être départagé au début de chaque épisode entre les régimes haussier et normal avec des probabilités respectives de l'ordre 48% et 41%. La situation devient de plus en plus déséquilibrée avec le temps au profit du premier régime sans pour autant que ce dernier ne soit considéré la seule alternative possible.

D'un autre coté, les coefficients η_i estimés pour les modèles (M2) et (M3), retenus pour mettre l'accent sur la relation rendement/risque, sont tous significativement différents de zéro. Le caractère significatif de ces coefficients justifie pleinement l'idée que la relation entre le rendement et le risque n'est ni stable sur le temps ni linéaire. En d'autres termes, ce *trade-off* rendement-risque varie considérablement à travers les différents états (Kim et Zumwalt, 1979). A ce propos, Guidolin et Timmermann (2004) rajoutent que la présence de ces différents régimes affecte significativement l'allocation optimale d'actifs détenus par les participants au marché boursier : En percevant une probabilité élevée d'être dans un état baissier, un investisseur détiendrait des proportions faibles dans des titres risqués sur le court terme. Cet investisseur détiendra plus d'actifs risqués à des horizons d'investissement plus long, comme la probabilité de changer à l'état haussier croît. A l'inverse, dans un état haussier, les investisseurs préfèrent détenir moins de titres risqués et plus d'actifs non risqués étant donné que la probabilité de changer à l'état baissier ou normal croît relativement à travers le temps.

Pour le besoin de la datation du marché boursier tunisien, nous avons calculé la variable qualitative It indiquant l'état dans lequel se trouve le marché à l'instant t. Dans un premier temps, nous devons affecter chaque observation à un régime donné. La règle de classification se présente comme suit :

$$I_t = \begin{cases} 1\ si\ P(S_t = 1\ /I_T) > P(S_t = 2\ /I_T)\ et\ P(S_t = 1\ /I_T) > P(S_t = 3\ /I_T) \\ 2\ si\ P(S_t = 2\ /I_T) > P(S_t = 1\ /I_T)\ et\ P(S_t = 2\ /I_T) >> P(S_t = 3\ /I_T) \\ 3\ si\ P(S_t = 3\ /I_T) > P(S_t = 1\ /I_T)\ et\ P(S_t = 3\ /I_T) > P(S_t = 2\ /I_T) \end{cases}$$

Les probabilités $P(S_t = i\ /I_T)$ (i = 1, 2, 3) représentent les probabilités non conditionnelles que le marché se trouve dans l'état i à l'instant t. Ces probabilités sont calculées à l'issu de l'estimation du modèle d'une manière récursive à partir de la densité des rendements construite comme un mélange de densités gaussiennes conditionnelles. En exploitant toute l'information disponible dans l'échantillon sur l'intervalle de temps [1, T], ces probabilités dites lissées sont préférées aux probabilités filtrées $P(S_t = i\ /I_t)$, qui elles se limitent seulement à l'information disponible jusqu'à l'instant t.

Dans un second temps, détecter les points de retournements entre les trois régimes en identifiant les pics et les creux à travers le calcul des indices de retournements (fig.5)[12]. Par la suite, la règle de classification entre pics et creux se présente comme suit :

[12] Indice de Retournement = 10* ($P[S_t = 1/I_t] - P[S_t = 3/I_t]$).

Pic à la date t si: 10* $(P[S_t = 1/I_t] - P[S_t = 3/I_t]) \geq 5$

Creux à la date t si: 10* $(P[S_t = 1/I_t] - P[S_t = 3/I_t] \leq -5$

Fig .4: Evolution hebdomadaire du rendement de l'indice TUNINDEX de 07/01/1998 au 19/08/2010

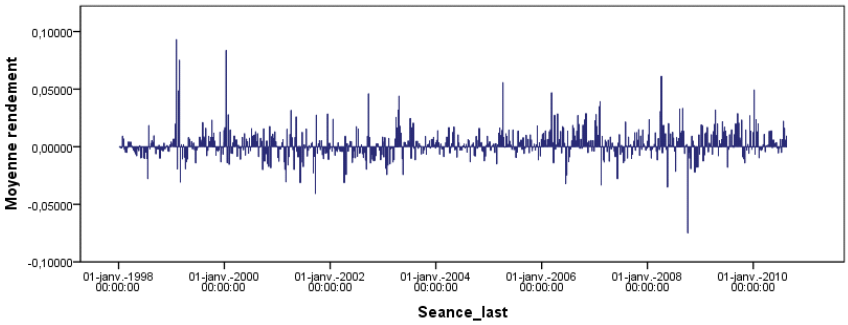
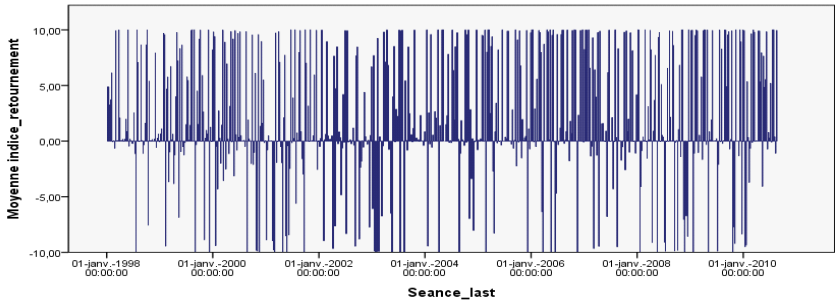


Fig .5: Indice de Retournement



		Numéro de l'observation	Indice de Retournement	
Rendement	*Bloc Supérieur*			
	1	58	9.59	**Régime 1**
	2	108	9.42	**Régime 1**
	3	61	5.91	**Régime 1**
	4	541	8.31	**Régime 1**
	5	384	9.40	**Régime 1**
	6	432	8.07	**Régime 1**
	7	**60**	**4.79**	**Régime 2**
	8	634	9.94	**Régime 1**
	9	280	6.96	**Régime 1**
	10	249	8.54	**Régime 1**
	11	432	8.07	**Régime 1**
	12	540	9.96	**Régime 1**
	13	172	7.21	**Régime 1**
	14	279	10.00	**Régime 1**
	15	595	9.99	**Régime 1**
	16	559	9.90	**Régime 1**
	17	562	9.88	**Régime 1**
	18	438	10.00	**Régime 1**
	19	617	9.52	**Régime 1**
	20	466	9.99	**Régime 1**
	21	540	9.96	**Régime 1**
	22	172	7.21	**Régime 1**
	23	279	10.00	**Régime 1**
	Bloc Inférieur			
	1	567	-9.78	**Régime 3**
	2	196	-10.00	**Régime 3**
	3	547	-9.96	**Régime 3**
	4	481	-1.18	**Régime2**
	5	446	-9.96	**Régime 3**
	6	225	-9.62	**Régime 3**
	7	181	-8.36	**Régime 3**
	8	62	-3.58	**Régime2**
	9	167	-9.96	**Régime 3**
	10	497	-6.57	**Régime 3**
	11	29	-9.99	**Régime 3**
	12	447	-4.54	**Régime2**
	13	268	-10.00	**Régime 3**
	14	284	-10.00	**Régime 3**

Tableau2 : Datation des points de retournement détectés par le modèle **DD(τ)-MS(3)-AR(2)-GARCH-M (1,1)**

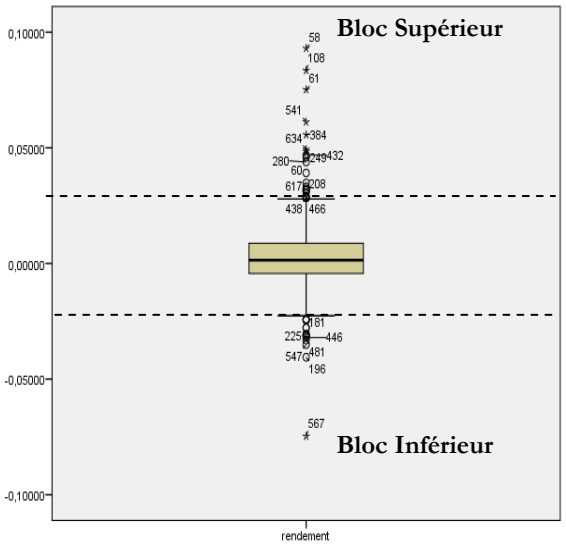


Figure 6 : Distribution des rendements hebdomadaires de l'indice TUNINDEX
(Semaines du 7 janvier 1998 au 19 août 2010)

D'après le tableau 2 et les figures 4, 5 et 6, nous pouvons avancer que les modèles à changements de régimes markoviens et tout particulièrement notre modèle DD(τ)-MS(3)-AR(2)-GARCH-M (1,1) peut être retenue comme un outil très pertinent en matière de détection et prédiction des points de retournement de l'activité boursière.

Conclusion :

Dans cette étude, le modèle proposé met en évidence le rôle important d'une modélisation non linéaire dans l'explication des dynamiques boursières sur le marché d'actions tunisien. En effet, les résultats montrent que le modèle à changement de régimes markovien à trois états offre plus de précision ; comparé à celui à deux états, particulièrement dans l'explication des caractéristiques cycliques non linéaires dans les rendements boursiers de l'indice TUNINDEX telles que la volatilité ou encore le retour à la moyenne. De plus, nous constatons que la notion de la dépendance à la durée introduite dans ce modèle permet de mieux capter cette structure non linéaire et donc appréhender davantage le comportement des rendements boursiers ; ainsi avoir la possibilité de faire de bonnes prévisions. D'un autre côté, l'étude dévoile une hétérogénéité substantielle du comportement des investisseurs sur le marché boursier tunisien durant les différents régimes. Par ailleurs, il s'est avéré que notre modèle à trois états n'arrive pas à expliciter convenablement les tendances secondaires, c'est pourquoi une étude future pourrait s'intéresser à examiner de près ces phénomènes et cela à l'aide d'un modèle à changement de régime markovien à quatre états.

Références

[1] Biscarri, J.G. et De Gracia, F.P. (2002) Bulls and Bears: Lessons from some European Countries, *http://repec.org/res2002/Gomez.pdf.*

[2] Chauvet, M. et Potter, S. (2000) Coincident and leading indicators of the stock market, *Journal of Empirical Finance, 7, 87-111.*

[3] Chen, S.S. (2007) Does monetary policy have asymmetric effects on stock returns? , *Journal of Money, Credit and Banking, 39, 667–688.*

[4] Chen, S-W. et Shen, C-H. (2007) Evidence of the Duration-Dependence from the Stock Markets in the Pacific Rim Economies, *Applied Economics, 39, 1461–1474.*

[5] Clark, P.K. (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices, *Econometrica, 41: pp. 135-155*

[6] Dueker, M. J. (1997) Markov Switching in GARCH Process and Mean Reverting Stock Market Volatility, *Journal of Business and Economic Statistics , 15, 26-34.*

[7] Durland, J.M. et McCurdy, T.H. (1994) Long Swings in the Dollar: Are they in the Data and Do the Market know it? *American Economic Review, 80, pp.689-713*

[8] Engle, R. F (1982) Autorgressive Conditional Heteroscedasticity with Estimates of U.K. Inflation, *Econometrice 50, 987-1008.*

[9] Fama, E. F. (1965) The Behavior of stock Market Prices, *Journal of Business 38, pp. 34-105.*

[10] Goldfeld, S.M. et Quandt, R.E. (1973) The estimation of structural shifts by switching regressions, *Annals of Economic and Social Measurement 2, 475-485.*

[11] Gonzalez, L; Hoang, P; Powell, J.G. et Shi, J. (2006) Defining and Dating Bull and Bear Markets: Two Centuries of Evidence, *Multinational Finance journal, 10, ½, 8I-116.*

[12] Granger, C.W.J. (2001) Spurious regressions in econometrics, in B. H. Baltagi (ed.), A Companion to Theoretical Econometrics, *Blackwell, Oxford, pp. 557—561*

[13] Guidolin, M. et Timmerman, A. (2005) Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns, *The Economic Journal, 115, 111-143.*

[14] Hamilton, J.D. (1989) A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle, *Econometrica, 57, 357-384.*

[15] Henry, O.T. (2009) Regime switching in the relationship between equity returns and short-term interest rates in the UK, *Journal of Banking and Finance, 33, 405–414.*

[16] Kim, M.K. et Zumwalt, J.K (1979) An Analysis of Risk in Bull and Bear Markets, *Journal of Financial and Quantitative Analysis, vol. XIV, N°5.*

[17] Maheu, J.M. et McCurdy, T.H. (2000) Identifying Bull and Bear Markets on the Stock Returns, *Journal of Business and Economic Statistics, vol 18, N°1, 100-112.*

[18] Maheu, J.M. ; McCurdy, T.H. et Song, Y. (2009) Extracting Bull and Bear Markets from Stock Returns, *http://www.rcfea.org/papers/2009/bayesian/maheu.pdf.*

[19] Maheu, J.M. ; McCurdy, T.H. et Song, Y. (2010) Components of bull and bear markets: Bull Corrections and Bear Rallies, *Working Paper 402, http://repec.economics.utoronto.ca/files/tecipa-402.pdf.*

[20] Mandelbrot, B.B (1963) The Variation of Certain Speculative Prices, *Journal of Business 36, 394-416.*

[21] Pagan, A.R. et Sossounov, K.A. (2003) A Simple Framework for Analyzing Bull and Bear Markets, *Journal of Applied Econometrics, 18, 23-46.*

[22] Pelagatti, M. (2003) DDMS-VAR for Ox: a software for time series modeling with duration dependent Markov-switching vector autoregressions, *st OxMetrics User Conference, London.*

Vers une approche probabiliste de la dépendance à la durée et de la datation du cycle boursier tunisien

Emira TORJMEN
Doctorante
Institut Supérieur de Gestion de Tunis
Adresse : 41 Rue de la Liberté – Cité Bouchoucha. Le Bardo – 2000– Tunis. Tunisie
Email : torjmen.emira@hotmail.fr

Dr. Adel KARAA
Professeur à l'Institut Supérieur de Gestion de Tunis
Adresse : 41 Rue de la Liberté – Cité Bouchoucha. Le Bardo– 2000 – Tunis. Tunisie
Email : adel.karaa@isg.rnu.tn ou adelkaraa@yahoo.fr

Introduction

La recherche d'une représentation adéquate de la dynamique sous-jacente aux rendements boursiers a fait l'objet de plusieurs études empiriques. La majorité de ces travaux s'est intéressée à la recherche des relations linéaires[1] entre les séries financières afin de décrire la dynamique de ces dernières. La simplicité des modèles linéaires pourrait être une explication du grand intérêt porté à cette spécification. Néanmoins, le recours à ces modèles conduit à une exploitation incomplète de l'information.

Avec l'accumulation des travaux sur les séries chronologiques macroéconomiques et financières, une panoplie de propriétés empiriques appelées aussi faits stylisés a été établie. Mandelbrot (1963) est le premier à avoir constaté deux faits caractéristiques, à savoir les queues de distribution plus épaisses que celle de la loi normale et les regroupements de la volatilité « *clustering effect* ». Depuis, d'autres faits stylisés[2] relatifs aux rendements des actifs financiers et à la volatilité ont été constatés. Ces différents faits empiriques conduisent à affirmer le caractère non-linéaire de la dynamique boursière, ainsi qu'une forte présence de changements structurels et de valeurs extrêmes.

Par ailleurs, une modélisation des séries nécessite le recours à des modèles spécifiques qui sont capables de reproduire le caractère non-linéaire. Afin de palier à ces insuffisances, plusieurs approches ont été proposées, les plus connues sont les spécifications non-linéaires de type ARCH (*AutoRegressive Conditional Heteroskdasticity*) introduites par Engle (1982). Ces processus visent principalement à rendre compte du fait que la variance conditionnelle n'est pas constante dans le temps.

Toutefois, les processus de type ARCH tiennent compte seulement du caractère non-linéaire. Ils exhibent leurs limites dans le cas où la volatilité des séries financières subit des changements de régimes occasionnels causés par certains événements, comme un krach boursier.

L'une des solutions qui fût apportée à ce problème a été d'envisager à ce que les paramètres d'une série temporelle puissent changer dans le temps, introduisant ainsi des changements paramétriques au sein des modèles non-linéaires. Partant de cette optique, une nouvelle classe de modélisation a été introduite par Hamilton (1989), à savoir les modèles à changements de régimes markoviens notés *MS* « *Markov Switching model*». L'approche de Hamilton part de l'idée que les séries chronologiques exhibent occasionnellement des coupures dans leur comportement avec une forte présence de valeurs aberrantes. Ces coupures sont associées à l'arrivée de l'information sur le marché boursier. Mandelbrot et Taylor (1967) et Clark (1973) ont démontré un lien entre les rendements des actifs et le taux de flux de l'information.

Dans le but d'expliquer ces ruptures, Hamilton (1989) met en évidence l'existence de plusieurs épisodes à travers lesquels la dynamique d'une série temporelle est différente. Il stipule que ces épisodes sont générés par des sauts discrets de régime au sein du processus sous-jacent à la dynamique de la série.

[1]Un processus linéaire chronologique est un modèle qui relie de façon linéaire la valeur présente d'une variable à ses valeurs passées, ainsi qu'à la valeur présente d'un bruit (Jawadi et Sahut, 2009).
[2] Cont (2001).

Maheu et McCurdy (2000) introduisent la dépendance à la durée[3] au sein des modèles à changement de régimes markoviens DDMS-DD (*Duration-Dependent Markov Switching Model*). Ils partent de l'idée que la probabilité que le marché boursier passe d'un état à un autre est variables dans le temps. D'où l'hypothèse de dépendance à la durée, qui n'est que la dépendance de l'intensité de changement à l'égard de l'âge et la maturité du régime. La prise en compte de cette notion permet de capter quelques structures non-linéaires au niveau de la moyenne conditionnelle et la variance conditionnelle dans les rendements boursiers américains. Le modèle DDMS-DD permet de détecter certains faits stylisés tels que le *clustering* de volatilité, le retour à la moyenne et les caractéristiques cycliques non linéaires dans les rendements.

Dans le même cadre d'analyse, l'étude de l'évolution de l'indice boursier TUNINDEX sur de longues périodes, nous a conduit à constater des ruptures brutales dans le comportement de ce dernier. La détection et la caractérisation de ces fluctuations cycliques s'inscrit en tant qu'une question primordiale afin d'évaluer et prévenir le comportement de l'indice sur une période donnée. En ce sens l'état du marché boursier tunisien varie dans le temps en passant par différentes phases distinctes. La détection de ces états de marché n'est pas visible à l'œil nu, la méthode à changements de régimes markovien se montre la méthode la plus adéquate.

Dans ce qui suit, nous proposons d'identifier et analyser les différentes phases cycliques caractérisant la dynamique du rendement de l'indice bousier TUNINDEX, de mettre en évidence le phénomène de dépendance à la durée, et d'établir une datation des cycles boursiers du marché tunisien. Notre étude porte sur une série de rentabilités hebdomadaires de l'indice boursier TUNINDEX sur une période s'étalant du 07/01/1998 au 19/08/2010.

1. Modèle

Nous développons une nouvelle approche plus réaliste de la variabilité des séries chronologiques financières en introduisant le modèle MS-AR-EGARCH. Ce modèle intègre à la fois les caractéristiques du modèle autorégressif à changements de régimes markovien de Hamilton (1989) et celles du modèle EGARCH à changements de régimes développé par T. Henry (2009). L'application de cette modélisation jointe de la moyenne et de la variance conditionnelle des séries étudiées, nous permet ainsi d'établir des fonctions de rentabilité et des fonctions de volatilité que nous pouvons utiliser afin d'étudier la dynamique cyclique dans le contexte du marché tunisien.

Le modèle MS(3) -AR(L) -EGARCH(1,1)[4] est définit comme suit:

$$R_t = \mu(S_t) + \sum_{q=1}^{L} \beta_q(S_t)[R_{t-q} - \mu(S_{t-q})] + \varepsilon_t \tag{1.1}$$

$$\varepsilon_t = \sqrt{h_t(S_t)}v_t \qquad v_t \to NID(0,1)$$

$$ln\big(h(s_{t-1})\big) = \delta_0(s_t) + \delta_1(s_t)\frac{\varepsilon_{t-1}}{h(S_{t-1})} + \delta_2(s_t)\left[\left|\frac{\varepsilon_{t-1}}{h(S_{t-1})}\right| - \sqrt{2/\pi}\right] + \delta_3(s_t)\ln\big(h(s_{t-1})\big) \tag{1.2}$$

avec ; $s_t = 1, 2, 3$;

R_t : le rendement actuel de l'indice TUNINDEX à l'instant t ;

$\mu(S_t)$: une constante dépendante du régime S_t ;

[3]Le modèle initié par Hamilton (1989) ne prend pas en considération la notion de la dépendance à la durée dans les états. Pour cela, Durland et McCurdy (1994) modifient la structure des probabilités de transition pour les permettre d'être dépendant de la durée (*Duration-Dependent Transition probabilities*).

[4] C'est une démarche statistique qui vient justifier la pertinence de la spécification retenue. Elle consiste à étudier dans une première étape la stationnarité de la série des rendements hebdomadaires de TUNINDEX moyennant les tests de Duckey-Fuller augmenté et de Perron. De déterminer dans une deuxième étape l'ordre des processus AR retenu pour décrire la dynamique des rendements moyennant les critères d'information AIC et BIC. Cette procédure est complétée par l'étude des fonctions d'auto-corrélation simple et partielle et des résultats du test Ljung-Box portant sur les données filtrées (i.e. la série des résidus issus de l'ajustement de la série des rendements par un modèle ARMA). Le test du Multiplicateur de Lagrange (LM) est utilisé dans une dernière étape pour tester la présence des effets ARCH dans la série des rendements filtrés. Le choix de trois régimes est justifié en faisant appel au test de rapport de vraisemblance adapté par Hansen (1992, 1996) et Garcia (1998) aux modèles à changements de régimes markovien.

Dans l'équation (1.1) l'évolution du rendement de l'indice du marché est supposée être régie par un processus autorégressif d'ordre L. La spécification log-linéaire (i.e. forme EGARCH) retenue dans l'équation (1.2) porte sur le logarithme de la variance conditionnelle afin d'éviter les contraintes de non-négativité des paramètres $\delta_0(.)$, $\delta_1(.)$ et $\delta_3(.)$ imposées dans les modèles GARCH. De plus, elle autorise à la variance conditionnelle de répondre de façon asymétrique au signe des innovations ε_{t-1} (ou encore les innovations standardisées : $\frac{\varepsilon_{t-1}}{h(S_{t-1})}$). Les effets de signe et d'amplitude dans chaque régime sont respectivement pris en compte par les coefficients $\delta_1(s_t)$ et $\delta_2(s_t)$. Dans ces conditions, ε_t peut être considérée comme une mesure globale traduisant l'information qui est arrivée sur le marché durant la période entre t-1 et t.

La dynamique de transition entre les états sont supposées être régie par un processus markovien d'ordre 1 avec les probabilités de transition suivantes :

$$P[S_t = j/S_{t-1} = i, S_{t-2} = k, \dots] = P[S_t = j/S_{t-1} = i] = P_{ij} \quad (1.3)$$

Pour chaque régime i (i = 1, 2, 3), les probabilités de transitions sont spécifiées de la manière suivante :

$$P_{ii} = P[S_t = i/S_{t-1} = i] = \frac{1}{1+\exp(\alpha_{ij})+\exp(\alpha_{ik})} \quad (1.4a)$$

$$P_{ij} = P[S_t = i/S_{t-1} = j] = \frac{\exp(\alpha_{ij})}{1+\exp(\alpha_{ij})+\exp(\alpha_{ik})} \quad (1.4b)$$

$$P_{ik} = P[S_t = i/S_{t-1} = k] = \frac{\exp(\alpha_{ik})}{1+\exp(\alpha_{ij})+\exp(\alpha_{ik})} \quad (1.4c)$$

La probabilité présentée en (1.4a) indique une situation de persistance du régime i (pas de changement de régime). Tandis que celles fournies en (1.4b) et (1.4c) expriment des probabilités de transition et traduisent respectivement les probabilités instantanées de quitter le régime i au profit des régimes j et k.

Les paramètres de l'équation du rendement et de la volatilité de l'indice de marché, ainsi que les probabilités de transition sont jointement estimés par la méthode de maximum de vraisemblance selon l'approche retenue par Hamilton (1989) et Kim (1994).

En s'inscrivant dans une optique cherchant à étudier la dépendance à la durée des probabilités de transition, nous avons fait appel à un modèle hybride DM-LM[5] selon lequel les probabilités de transition relatives à chaque régime i (i = 1, 2, 3) peuvent s'exprimer comme suit :

$$P_{ii}(d_{t-1}) = P(I_t = i/I_{t-1} = i, d_{t-1}) = \frac{\alpha_{ii}}{\alpha_{ii} + \alpha_{ij}\exp(\beta_{ij}d_{t-1}) + \alpha_{ik}\exp(\beta_{ik}d_{t-1})}$$

$$P_{ij}(d_{t-1}) = P(I_t = j/I_{t-1} = i, d_{t-1}) = \frac{\alpha_{ij}\exp(\beta_{ij}d_{t-1})}{\alpha_{ii} + \alpha_{ij}\exp(\beta_{ij}d_{t-1}) + \alpha_{ik}\exp(\beta_{ik}d_{t-1})}$$

$$P_{ik}(d_{t-1}) = P(I_t = k/I_{t-1} = i, d_{t-1}) = \frac{\alpha_{ik}\exp(\beta_{ik}d_{t-1})}{\alpha_{ii} + \alpha_{ij}\exp(\beta_{ij}d_{t-1}) + \alpha_{ik}\exp(\beta_{ik}d_{t-1})}$$

avec I_t une variable polytomique indiquant l'état dans lequel se trouve le marché à l'instant t :

$$I_t = \begin{cases} 1\ si\ P(S_t = 1/I_T) > P(S_t = 2/I_T)\ et\ P(S_t = 1/I_T) > P(S_t = 3/I_T); \\ 2\ si\ P(S_t = 2/I_T) > P(S_t = 1/I_T) et\ P(S_t = 2/I_T) > P(S_t = 3/I_T); \\ 3\ siP(S_t = 3/I_T) > P(S_t = 1/I_T) et\ P(S_t = 3/I_T) > P(S_t = 2/I_T). \end{cases}$$

$P(S_t = i/I_T)$, (i = 1, 2, 3) étant la probabilité non conditionnelle que le marché se trouve dans l'état i à l'instant t. Ces probabilités sont calculées à l'issu de l'estimation du modèle de manière récursive à partir de la densité des rendements construite comme un mélange de densités gaussiennes conditionnelles. En exploitant toute l'information disponible dans l'échantillon sur l'intervalle de temps [1, T], ces probabilités dites lissées sont préférées aux probabilités filtrées $P(S_t = i/I_t)$, qui elles se limitent seulement à l'information disponible jusqu'à l'instant t.

Chaque paramètre α_{ij}, (i, j = 1,2,3) indique l'importance de chaque transition du type (i,j). Ces derniers peuvent aussi être considérés comme des facteurs d'hétérogénéité non observables. En effet, une valeur élevée (faible) du paramètre, indique une faible (forte) hétérogénéité. Une statistique

[5] Pour plus de détail sur ce type de modèle, on peut se référer à Fader 1993.

souvent utilisée afin de détecter le niveau global d'une hétérogénéité qui est l'indice de polarisation «*Polarization index*»[6] exprimé comme suit :

$$\varphi = \frac{1}{1+S}$$

avec $S = \sum_{i,j} \alpha_{ij}$ et i, j=1, 2 ,3.

L'indice de polarisation est une mesure d'hétérogénéité relative à la présence discrète et répétée du marché dans un régime donné. Cet indice prend sa valeur entre 0 et 1. Si l'indice est très proche de 0, nous pouvons déduire une très faible hétérogénéité inversement ; si la valeur de l'indice est très proche de 1, nous pouvons déduire une très forte hétérogénéité.

Au niveau de chaque régime i, on parle d'une dépendance des probabilités de transitions vis-à-vis l'ancienneté du régime lorsqu'au moins un des coefficients β_{ij} , $(i, j = 1,2,3)$ est significativement différent de zéro. Cette dépendance est dite positive (respectivement, négative) si de plus ces coefficients sont positifs (respectivement, négatifs). Dans ces conditions, les probabilités de transition se montrent comme croissantes (respectivement, décroissantes) en fonction de l'ancienneté du régime *i*, auquel elles associent par la même occasion un niveau de persistance qui est décroissant (respectivement, croissant) au fil du temps.

La fonction de vraisemblance du modèle est définie comme suit :

$$V = \prod_{t=1}^{T}\left[\sum_{i=1}^{3} h_i(t-1)\left[\prod_{j=1}^{3}\left(P_{ij}\,(d_{t-1}\right)^{h_j(t)}\right]\right]$$

avec $h_j(t) = \begin{cases} 0 & si\ I_t = j \\ 1 & sinon \end{cases}$

Ainsi établie, la variable I_t nous a donné aussi les moyens de procéder à une datation du cycle boursier tunisien. En annexe, nous avons pris le soin de vérifier la pertinence d'un tel processus en se référant à des dates bien précises ayant marquées l'évolution des rendements de l'indice TUNINDEX.

2. Résultats empiriques

Les différents résultats dont nous avons aboutit à partir de l'application du modèle à changements de régimes markovien à trois états sur les rendements hebdomadaires de l'indice boursier TUNINDEX sont exposés dans le tableau 2.1.

Tableau 2.1: Présentation des résultats des estimations du modèle MS(3)-AR(2)-EGARCH(1,1) par la méthode du maximum de vraisemblance

Equation de rendement	μ_1	-1.4589 (0.0220) ***
	μ_2	0.6756 (0.0068) ***
	μ_3	3.8678 (0.0469) ***
	β_{11}	-0.3308 (0.0344) ***
	β_{12}	0.1565 (0.0028) ***
	β_{21}	0.3956 (0.0314) ***
	β_{22}	0.1474 (0.0112) ***
	β_{31}	0.2398 (0.0026) ***
	β_{32}	0.8961 (0.0186) ***
Equation de volatilité	δ_{01}	0.2236 (0.1189) **
	δ_{02}	-3.9649 (0.0710) ***
	δ_{03}	-0.2424 (0.1543) *
	δ_{11}	0.2757 (0.0353) ***
	δ_{12}	-0.3446 (0.0499) ***
	δ_{13}	0.0083 (0.0496)
	δ_{21}	0.4904 (0.0443) ***
	δ_{22}	-0.7474 (0.1004) ***
	δ_{23}	0.5935 (0.0703) ***
	δ_{31}	-0.0404 (0.0289) **
	δ_{32}	0.6231 (0.0166) ***
	δ_{33}	0.0698 (0.0293) ***
Probabilité de transition	α_{12}	3.6244 (0.2502) ***
	α_{13}	0.6763 (0.2442) ***
	α_{21}	-6.3930 (0.3603) ***
	α_{23}	-9.1650 (0.5559) ***
	α_{31}	-0.3938 (0.1435) ***
	α_{32}	2.8173 (0.2173) ***
Log-vraisemblance		*-6,5579022*
Nombre d'observations		*666*

Notes : - * présentent un niveau de signification de 10%, ** présentent un niveau de signification de 5%,*** présentent un niveau de signification de 1%.
- Les valeurs indiquées entre parenthèses représentent les écarts types.

D'après le tableau 2.1, nous pouvons remarquer que les moyennes conditionnelles $\mu(S_t)$ attachées aux trois régimes du modèle sont tous significatives avec ;

[6] Fader (1993).

$$\tilde{\mu}(1) = (-1.4589) < \tilde{\mu}(2) = (0.6756) < \tilde{\mu}(3) = (3.8678)$$

Un premier régime avec une moyenne de rendement estimée de (-1.4589), un deuxième régime présentant une moyenne de rendement estimée de (0.6756) et enfin un troisième régime avec une moyenne de rendement estimée de (3.8678).

En ce qui concerne l'analyse de la volatilité, le premier et le troisième régime sont caractérisés par une grande volatilité inconditionnelle (0.2231, -0.0242) par rapport à celle définie dans le deuxième régime (-3.9649).

L'impact de l'arrivée des nouvelles pendant la période t sur le niveau de volatilité est capté par le coefficient $\delta_2(S_t)$ de l'équation (1.2). Nous pouvons remarquer, d'après le tableau 2.1 que $\widehat{\delta_{21}} > \widehat{\delta_{23}} > \widehat{\delta_{22}}$, ce qui implique que les nouvelles qui arrivent lorsque le marché se trouve dans le régime 1 ou dans le régime 3 conduisent à une augmentation de volatilité relativement importante par rapport au régime 2.

Le coefficient $\delta_3(S_t)$, nous renseigne sur le niveau de persistance des chocs au niveau de la volatilité conditionnelle. Au vu des résultats obtenus, nous pouvons détecter un niveau de persistance très faible dans le régime 1 et le régime 3 de l'ordre de (-0.0404) et (-0.0698) respectivement pour les deux régimes, contrairement au régime 2 où nous enregistrons une forte persistance des chocs de l'ordre de 0.6231.

$$\widehat{\delta_{32}} > \widehat{\delta_{31}} > \widehat{\delta_{33}}$$

En tenant compte du sens de l'information, nous pouvons déduire à partir des résultats obtenus, qu'une bonne nouvelle introduite sur le marché étant dans lé régime 2 ($\varepsilon_{t-1} > 0$), conduira à diminuer la volatilité puisque $\widehat{\delta_{12}} < 0, \widehat{\delta_{22}} < 0$. Quant à l'arrivée d'une mauvaise nouvelle($\varepsilon_{t-1} < 0$), elle sera traduite par une hausse de la volatilité.

Dans le régime 1 et le régime 3, le marché semble afficher une autre réponse à l'arrivée des nouvelles. En effet, une bonne nouvelle ($\varepsilon_{t-1} > 0$) tend à accroître la volatilité sur le régime 1 étant donnée que $\widehat{\delta_{11}} > 0, \widehat{\delta_{21}} > 0$. La volatilité au niveau du régime 3, elle va augmenter avec l'introduction d'une bonne nouvelle sur le marché puisque $\widehat{\delta_{13}} > 0, \widehat{\delta_{23}} > 0$. Quant à l'arrivée d'une mauvaise nouvelle ($\varepsilon_{t-1} < 0$) , la volatilité tend à diminuer au niveau du régime 1 et du régime 3.

Ces constations concourent à retenir le caractère volatile[7] des rendements au niveau des deux régimes 1 et 3. Dès lors, ces derniers seront sensibles aux bonnes et aux mauvaises nouvelles contrairement au deuxième régime caractérisé par des rendements non-volatils.

En ce qui concerne les probabilités de transition et les probabilités de persistance, ils sont exposés dans le tableau 2.2.

Tableau 2.2 : Matrice de transition estimée

	Régime 1	**Régime 2**	**Régime 3**
Régime 1	p_{11} : 0.0247	p_{12} : 0.9266	p_{13} : 0.0485
Régime 2	p_{21} : 0.0016	p_{22} : 0.9982	p_{23} : 0.0001
Régime 3	p_{31} : 0.0366	p_{32} : 0.9090	p_{33} : 0.0054

Il y a lieu de capter la valeur la plus élevée de la matrice qui est de l'ordre de 0.9982. Cette valeur est respectivement liée à la probabilité P_{22} qui nous permet de se rendre compte du caractère relativement persistant du régime 2. En revanche, les valeurs estimées de 0.0247 et 0.0054 respectivement pour les probabilités P_{11}et P_{33} ne peuvent que révéler le caractère extrêmement transitoire des régimes 1et 2. Nous pouvons aussi enregistrer une forte probabilité de quitter le régime 1 au profit du régime 2 de l'ordre de 93% (P_{12}=0.9266), de même, la probabilité du passage du marché du régime 3 au régime 2 est trop élevée avec une valeur près de 91% (P_{32}=0.9090). Au vu de ces résultats, nous pouvons constater l'effet attractif du régime 2. Autrement dit, le marché tend à converger vers le régime 2, aux dépens des deux autres régimes.

A partir des résultats relatifs aux probabilités non conditionnelles (probabilités lissées) exposés au niveau du tableau 2.3, nous pouvons constater que le régime 2 présente la valeur la plus

[7] Plus une action sera « volatile » et plus son cours sera sensible aux bonnes et aux mauvaises nouvelles concernant l'entreprise ou les marchés.

élevée par rapport aux deux autres régimes. Autrement dit, le régime est le plus présent dans le marché boursier tunisien, il présente la fréquence la plus élevée.

Tableau 2.3 : Probabilités non conditionnelles

Régime 1	Régime 2	Régime 3
0,1989	0,5326	0,2683

Les constations tirées des différents résultats obtenus mettent en avant l'existence de trois régimes distincts pour décrire l'évolution des cours du TUNINDEX dans le temps.

Au vu des différentes constations, nous pouvons associer les deux régimes 1 et 3 caractérisés par un niveau relativement élevé de volatilité et un niveau faible de persistance aux régimes relatifs aux périodes d'événements (régimes événementiels). En effet, Brown et Warner (1985) et Boehmer et *al*[8] (1991) dans le cadre de leurs études portées sur l'hétérovariance liée aux événements, ils ont constaté un accroissement de la variance lié aux périodes événementielles. Autrement dit, la variance a tendance à être plus élevée pendant la période d'événement par rapport à la période hors événement. Toutefois, les deux régimes présentent une divergence au niveau du rendement. Zhang (2006) a détecté un lien entre le sens de l'information et le niveau du rendement futur. En effet, les résultats de son étude ont montré qu'à un niveau élevé d'incertitude de l'information, les rendements futurs seront relativement bas (élevé) suivant les mauvaises (bonnes) nouvelles. Par ailleurs, le niveau faible du rendement dans le régime 1 nous permet d'associer ce dernier à un régime événementiel de mauvaises nouvelles. En ce qui concerne le régime 3, caractérisé par un niveau élevé de rendement, nous pouvons l'associer à un régime événementiel de bonnes nouvelles. Finalement, le niveau de persistance relativement élevé, le niveau relativement faible de la volatilité et le rendement proche de zéro enregistrés dans le régime 2, nous donne le droit de l'associer à un régime de hors événement. Le recours aux probabilités instantanées non conditionnelles ne peut que confirmer nos propos. En effet, ces dernières montrent que la fréquence d'apparition du régime 2 est plus élevée que la fréquence des deux autres régimes, ce qui nous permet de qualifier ce dernier comme un régime habituel. Ainsi, le marché boursier tunisien tend le plus à rejoindre le régime hors événements chaque fois qu'il se trouve soit dans ce même régime ou dans un autre régime. Les caractéristiques de chaque régime sont présentées dans le tableau 2.4.

	Rendement	Volatilité	Persistance dans le régime	Fréquence
Régime 1	Faible	Elevé	Faible	Faible
Régime 2	≈ nul	Faible	Elevée	Elevée
Régime 3	Elevé	Elevé	Faible	Faible

Tableau 2.4 : Caractéristiques des trois régimes (rentabilité, volatilité, persistance, fréquence)

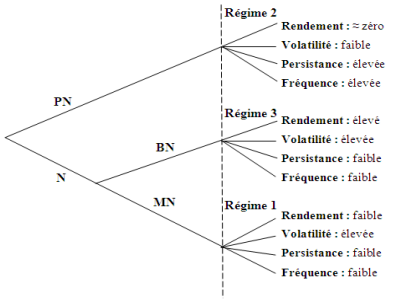


Fig. 2.1 : Effet de l'annonce sur l'évolution de l'indice TUNINDEX[9]

Au vu des différentes constations, nous pouvons qualifier l'évolution du processus sous-jacent à la dynamique des rendements de l'indice boursier TUNINDEX comme une alternance entre trois différents régimes. Cette alternance est provoquée par l'arrivée de nouveaux événements sur le marché. En effet, le marché passe d'un état à un autre suivant l'arrivée des événements sur le marché, il existe dans ce cadre trois situations possibles (pas de nouvelles, l'arrivée de bonnes nouvelles, l'arrivée de mauvaises nouvelles) ce qui explique nettement le nombre de régime choisit dans le cadre de notre étude (Fig. 2.1).

[8] Boehmer, Musumeci et Poulsen (1991).

[9] PN : Pas de nouvelle ; N : Annonce de nouvelle ; BN : Annonce de bonne nouvelle ; MN : Annonce de mauvaise nouvelle.

Nous allons, dans ce qui suit, mettre en évidence les résultats relatifs au phénomène de dépendance à la durée. Le tableau 2.5, correspondent aux coefficients estimés des probabilités de transition variables dans le temps du modèle DM-LM par la méthode du maximum de vraisemblance.

Tableau 2.5 : Présentation des résultats des estimations des probabilités de transition du modèle DM-LM par la méthode du maximum de vraisemblance

Probabilité de transition variable	α_{11}	0.2274 (0.022) ***
	α_{12}	0.5777 (0.1516) ***
	α_{13}	0.1806 (0.0394) ***
	α_{21}	0.1664 (0.0124) ***
	α_{22}	0.2342 (0.0120) ***
	α_{23}	0.1303 (0.0100) ***
	α_{31}	0.4400 (0.2022) **
	α_{32}	1.1002 (0.5336) **
	α_{33}	0.1757 (0.0241) ***
	β_{12}	0.6052 (0.1915) ***
	β_{13}	0.3744 (0.1607) ***
	β_{21}	-0.0831 (0.0168) ***
	β_{23}	-0.0569 (0.0177) ***
	β_{31}	0.9380 (0.3867) ***
	β_{32}	1.1510 (0.4000) ***
Log-vraisemblance		*-3812,469048*
Nombre d'observations		*666*

Notes : - * présentent un niveau de signification de 10%,, ** présentent un niveau de signification de 5%, *** présentent un niveau de signification de 1%.
- Les valeurs indiquées entre parenthèses représentent les écarts types.

Les coefficients, β_{ij} et β_{ik}, retenus en vu d'estimer la dépendance des probabilités de transition en terme de durée sont tous, d'après le tableau 3.11, statistiquement significatifs. Dans ce contexte, nous pouvons remarquer que les coefficients β_{21} et β_{23} sont tous les deux négatifs, respectivement de l'ordre de -0.0831 et -0.0.569, ce qui implique des fonctions de hasard décroissantes associées à ce cas. En d'autres termes, les deux probabilités de quitter le régime 2 au profit du régime respectivement 1 et 3 diminuent avec la durée. Dans ces conditions, le niveau de persistance s'avère être croissant au fil du temps.

En revanche, les coefficients notés ; β_{12} ; β_{13} et β_{31} ; β_{32} semblent être tous positifs qui sont de l'ordre respectivement de 0.6052 ; 0.3744 et 0.9380 ;1.1510. Les fonctions de hasard sont croissantes avec le régime 1 et le régime 2 ce qui atteste, par la même occasion, que la probabilité de quitter l'un ou l'autre régime augmente avec la durée. Ainsi, les deux régimes 1 et 3 présentent un niveau de persistance relativement faible par rapport au régime 2.

Les liens entre la durée et les phases cycliques de l'indice TUNINDEX sont mis en évidence par un ensemble de représentations graphiques retraçant l'évolution des probabilités de persistance et de transition en fonction de la durée qui sont présentées dans la fig. 2.2.

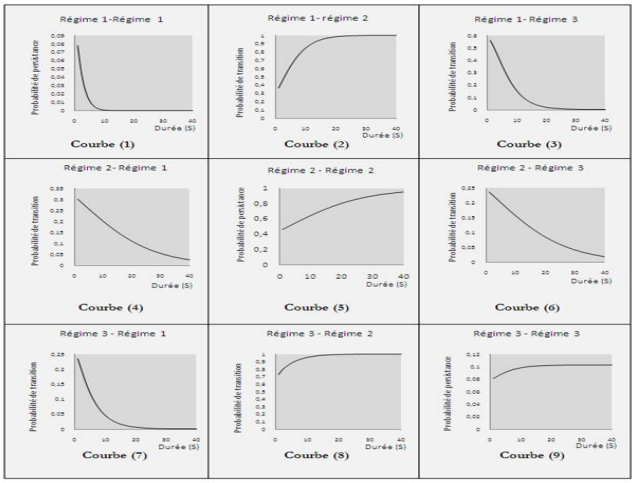


Fig. 2.2 : Evolution des probabilités de transition et de persistance en fonction de la durée

D'après les graphiques, nous pouvons constater que l'évolution des probabilités de transition et de persistance présente une forte dépendance à la durée. En effet, les courbes retraçant l'évolution des probabilités dans le temps peuvent être croissantes ou décroissantes suivant le régime étudié. Cette constatation rejoint les résultats établis par Maheu et McCurdy (2000) prouvant l'existence de dépendance à la durée au niveau des marchés haussiers et baissiers.

Nous pouvons remarquer à partir de la courbe (1), une forte décroissance de la probabilité de persistance du régime 1 de mauvaises nouvelles au fur et à mesure que la durée augmente. Cette faible persistance du régime 1 dans le temps avance l'idée de la forte présence de creux[10] dans l'évolution de l'indice boursier dans le temps.

En ce qui concerne la probabilité de persistance du marché au niveau du régime 3 de bonnes nouvelles dans le temps présentée par la courbe (9), nous pouvons constater qu'elle exhibe une forme croissante. Néanmoins, il faut noter que cette persistance est très faible avec un maximum enregistré à partir de la $14^{ème}$ semaine d'environ 0.1. Ce qui nous laisse penser que l'évolution du marché tunisien dans le temps est caractérisée par une présence de pics[11]. Cette présence de pics est plus ou moins inférieure à celle des creux.

Les courbes relatives au passage du régime 1 au régime 2 (courbe(2)) et du régime 3 au régime 2 (courbe(8)) présentent une forme croissante. Ceci signifie que les probabilités de transition dans ces deux courbes augmentent au fur et à mesure que la durée augmente. Cet accroissement atteint son maximum (≈ 1) au bout de ≈ 10 semaines. Ainsi, le régime 2 exerce un effet d'attraction dans le sens où les deux autres régimes ont tendance à converger dans un temps moindre vers ce dernier.

Au vu de ces constations, nous pouvons qualifier le régime hors événement (régime 2) d'une phase habituel contrairement aux deux régimes événementiels (régime 1, régime 3) qui représentent une phase transitoire. La dynamique du marché agit considérablement dans le temps pour ramener le marché à rejoindre le régime 2. Ce qui nous laisse introduire l'idée que le régime 2 est un régime d'équilibre, indiquant que le marché tunisien, en absence de nouvelles informations, il se corrige en agissant sur les différents écarts des cours par rapport à l'équilibre dus aux informations antérieurement annoncées.

Pour affermir davantage le diagnostic, nous allons porter une brève attention sur l'indice de polarisation. D'après les résultats obtenus dans le tableau 3.11, nous avons les valeurs de l'indice de polarisation pour chaque régime ;

$$(\alpha_{11}, \alpha_{12}, \alpha_{13}) \sim (0.2274, 0.5777, 0.1806)$$
$$\varphi_1 = 0.5036$$
$$(\alpha_{22}, \alpha_{21}, \alpha_{23}) \sim (0.2342, 0.1664, 0.1303)$$
$$\varphi_2 = 0.6532$$
$$(\alpha_{33}, \alpha_{32}, \alpha_{31}) \sim (0.1757, 1.1002, 0.4400)$$
$$\varphi_3 = 0.3682$$
$$\varphi_3 < \varphi_1 < \varphi_2$$

Les valeurs de l'indice de polarisation obtenues viennent confirmer nos constations, dans le sens où le régime 2 exhibe la valeur la plus élevée de l'indice. Ce niveau relativement élevé indique que le régime 2 présente un niveau élevé d'hétérogénéité aux dépens des autres régimes. Par ailleurs, nous pouvons introduire l'idée que le régime 2 exerce un effet de polarisation (d'attraction) sur les autres régimes. Autrement dit, les deux régimes événementiels relativement de bonnes et de mauvaises nouvelles ont tendance à converger vers un régime d'équilibre (régime hors événements).

Cette persistance au niveau du régime d'équilibre (régime 2) au fils du temps, peut être associée à la présence de retour à la moyenne[12] « *Mean Reversion* » des cours de l'indice. En effet,

[10] Les creux indiquent un passage du marché d'une période baissière suite à l'annonce d'une mauvaise nouvelle vers une autre période (période sans nouvelle ou période de bonne nouvelle).

[11] Les pics indiquent la fin de la période haussière suite à l'annonce d'une bonne nouvelle et que le marché va passer à une autre période (période sans nouvelle ou période de mauvaise nouvelle).

[12] Le phénomène de retour à la moyenne « *Mean reversion* » n'est qu'une conséquence de la présence de mémoire longue dans une série de rendements boursiers. Il implique que l'effet d'un choc sur la série aura un impact durable mais non permanent. Cet effet de persistance indique la durée du cycle engendrée par le choc.

dans un régime événementiel, les prix d'actions s'éloignent de leurs valeurs d'équilibre. Plus cette déviation des cours est importante plus le processus d'ajustement est actif et le retour des cours à l'équilibre est rapide. Ce qui explique la rapidité associée à l'ajustement des cours boursiers à l'équilibre. Cette rapidité d'ajustement explique la persistance du deuxième régime aux dépens de deux autres.

Conclusion

Dans le but d'identifier et d'analyser les différentes phases cycliques qui caractérisent la dynamique du marché boursier tunisien, nous avons eu recours aux modèles à changements de régimes markoviens. Notre étude s'est portée sur une série de rentabilités hebdomadaires de l'indice boursier TUNINDEX sur une période s'étalant du 07/01/1998 au 19/08/2010. La modélisation de cette série est représentée par un modèle MS-AR [Hamilton (1989)] pour les rendements et un modèle MS-EGARCH [T. Henry (2009)] pour la volatilité.

Les résultats obtenus de notre modèle mettent en évidence la présence de trois régimes distincts relatifs aux types des événements annoncés sur le marché, à savoir; un régime hors événements, un régime événementiel de bonnes nouvelles et un régime événementiel de mauvaises nouvelles. Une lecture probabiliste des résultats nous a conduits à affirmer le caractère persistant du régime hors événements en dépit des deux autres régimes.

En ce sens, nous sommes parvenus aussi à vérifier que la probabilité de sortir d'un régime dépend de la durée de celui-ci. En effet, l'examen des courbes d'évolution des probabilités de persistance et de transition en fonction de la duré nous a permis de conclure que la probabilité de persister dans le régime hors événements augmente avec le temps. La dynamique du marché agit considérablement dans le temps pour ramener le marché à rejoindre le régime hors événements. Ce qui nous laisse introduire l'idée que ce dernier est un régime d'équilibre, indiquant que le marché tunisien, en absence de nouvelles informations, il se corrige en agissant sur les différents écarts des cours par rapport à l'équilibre dus aux informations antérieurement annoncées. Cette persistance au niveau du régime d'équilibre au fils du temps, peut être associée à la présence de retour à la moyenne des cours de l'indice.

La mise en place d'une datation des fluctuations cycliques des rendements de l'indice indique que le modèle à changements de régimes markoviens, peut être considéré comme un outil fiable et intéressant pour la détection et la prédiction des points de retournement de l'activité boursière. En effet, les différents points de retournement enregistrés épousent les fluctuations extrêmes des rendements. Ce qui nous ramène à conclure que le modèle employé reproduit dans le cadre de notre étude reproduit au mieux la dynamique boursière des rendements de l'indice TUNINDEX.

Annexe 1 : Datation des points de retournements

La règle de classification, consiste à attribuer à l'observation à la date t le régime qui présente la probabilité lissée la plus élevée:

Si $P[S_t = 1/I_T] > P[S_t = 2/I_T]\ et\ P[S_t = 1/I_T] > P[S_t = 3/I_T]$; le régime 1 est assigné à l'observation à l'instant t.

Si $P[S_t = 2/I_T] > P[S_t = 1/I_T]\ et\ P[S_t = 2/I_T] > P[S_t = 3/I_T]$; le régime 2 est attribué à l'observation à l'instant t.

Si $P[S_t = 3/I_T] > P[S_t = 1/I_T] et\ P[S_t = 3/I_T] > P[S_t = 2/I_T]$; le régime 3 est assigné à l'observation à l'instant t.

D'une manière générale, la règle se présente comme suit ;

$$i^* = \max P[S_t = i/I_T]$$

Dans cette perspective, et après avoir attribué à chaque observation le régime qui lui correspond, nous allons dans ce qui suit essayer de détecter les points de retournements entre les trois régimes et d'établir une datation des points en identifiant les pics et les creux.

Les points de retournement présentés dans la figure 3.4 peuvent être détectés comme suit :

Cette durée de cycle n'est autre que le temps nécessaire pour que la série revienne à son niveau d'équilibre, autrement dit, le temps durant lequel la série diverge de cette valeur d'équilibre. L'arrivée d'un nouveau choc, conduit à l'apparition d'un autre cycle.

Retournement =10* ($P[S_t = 3/I_T] - P[S_t = 1/I_T]$)

La procédure de datation des points de retournement, pour le cas de trois régimes, s'effectue de la manière suivante (Fig. I.1):

Pic à la date t si: 10* ($P[S_t = 3/I_t] - P[S_t = 1/I_t]$) ≥ 5
Creux à la date t si: 10* ($P[S_t = 3/I_t] - P[S_t = 1/I_t] \leq -5$

Fig. I.2 : Evolution hebdomadaire du rendement de l'indice TUNINDEX de 07/01/1998 au 19/08/2010

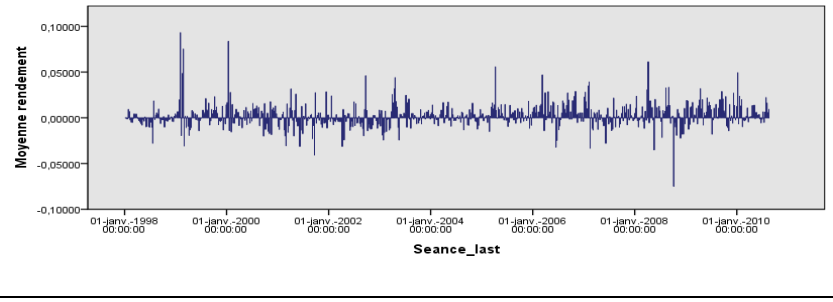
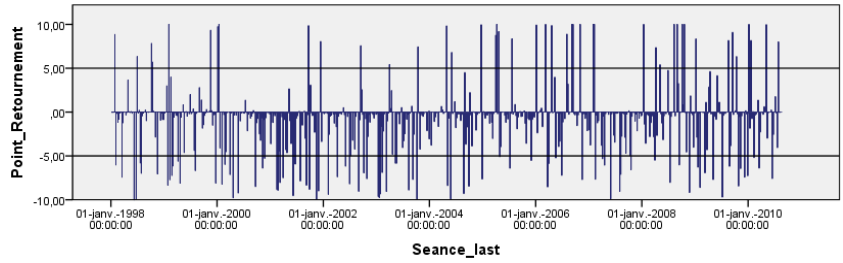


Fig. I.1: Point de Retournement



Afin de mieux cerner l'apport de notre travail, nous présentons dans le tableau I.1 un rapprochement entre les points de retournement enregistrés à travers le modèle à changements de régimes markoviens MS(3)-AR(2)-EGARCH(1,1) et les valeurs extrêmes caractérisant la dynamique des rendements de l'indice TUNINDEX enregistrés au niveau du bloc supérieur et bloc inférieur sur la période d'échantillonnage dans le figure de *Boxsplot* (Fig. I.3).

		Numéro de l'observation	Indice de retournement	Régime
Rendement	Bloc Supérieur			
	1	58	10	Régime 3
	2	108	10	Régime 3
	3	61	7,22	Régime 3
	4	541	7,35	Régime 3
	5	384	10	Régime 3
	6	634	10	Régime 3
	7	60	3,97	Régime 2
	8	432	10	Régime 3
	9	249	2,55	Régime 2
	10	280	5,80	Régime 3
	11	480	9,99	Régime 3
	12	479	10	Régime 3
	13	562	10	Régime 3
	14	595	4,63	Régime 3
	15	279	2,44	Régime 2
	16	540	5,49	Régime 3
	17	172	0,87	Régime 2
	18	466	10	Régime 3
	191	617	9,09	Régime 3
	20	438	9,87	Régime 3
	21	208	8,04	Régime 3
	Bloc Inférieur			
	1	567	-10	Régime 1
	2	196	-9,85	Régime 1
	3	481	-7,65	Régime 1
	4	547	-3,25	Régime 2
	5	225	-7,37	Régime 1
	6	446	-3,33	Régime 2
	7	181	-9,25	Régime 1
	8	497	-6,21	Régime 1
	9	29	-5,80	Régime 1
	7	447	-8,17	Régime 1
	8	268	-9,32	Régime 1
	9	284	-5,83	Régime 1
	10	227	-4,15	Régime 2

Tableau I.1 : Datation des points de retournement détectés par le modèle MS(3)-AR(2)-EGARCH(1,1)

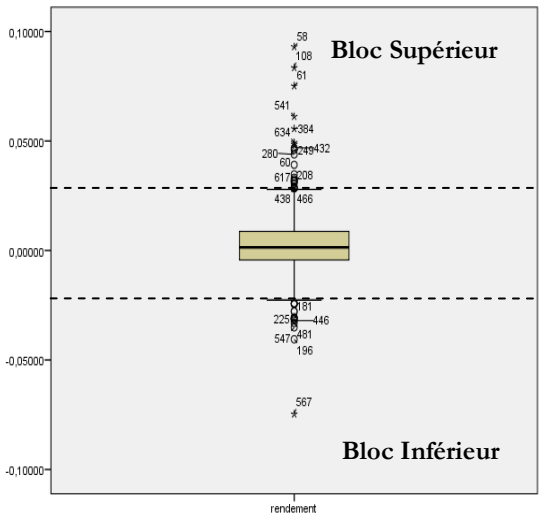


Fig. I.3 : Distribution des rendements hebdomadaires de l'indice TUNINDEX (Semaines du 7 janvier 1998 au 19 août 2010)

Nous pouvons qualifier les points de retournements les plus élevés (faible) dans le bloc supérieur (inférieur) de pics (creux) prononcés qui sont dus généralement à des événements phares sur le marché financier. A titre d'exemple, le creux au niveau de l'observation 196 (un indice de retournement≈10) enregistrée à la date du 21 septembre 2001 est une conséquence directe du replanissement économique connu suite aux attentats du 11 septembre 2001. Une grande période de récession économique a été connue à l'échelle mondial ce qui a affecté l'activité boursière à travers le monde entier.

A partir de ces constations, la méthode paramétrique des modèles à changements de régimes markoviens, spécifiquement le modèle MS(3)-AR(2)-EGARCH(1,1), peut être considérée comme un outil intéressant pour la détection et la prédiction des points de retournement de l'activité boursière. Cette méthode a déjà montré ses preuves au niveau de l'activité économique tunisienne à travers plusieurs travaux empiriques. Néanmoins, sur le marché boursier tunisien, notre travail est considéré comme une première. En effet, la dynamique cyclique du marché boursier tunisien n'ont pas fait l'objet d'une identification et une datation des points de retournements.

Références

[1] Brown, S.J. et J.B. Warner, (1985), "Using Daily Stock Returns: The Case of EventStudies". *Journal of Financial Economics 14, 3-31.*

[2] Boehmer, E., J. Musumeci, et A. Poulsen, (1991)"Event-Study Methodology Under Conditions of Event-Induced Variance". *Journal of Financial Economics 30, 253-272.*

[3] Clark, P., (1973), "A subordinated stochastic process model with finite variance for speculative prices". *Econometrica, 41, 135-156.*

[4] Cont, R., (2001), Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative Finance 1, 223–236.*

[5] Durland, J. M., et T. H. McCurdy., (1994), "Duration-Dependent Transitions in a Markov Model of U.S. GNP Growth''. *Journal of Business & Economic Statistics, 12(3), 279–288.*

[6] Engle, R., (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation". *Econometrica 50, 987-1007.*

[7] Fader, P., (1993). "Integrating the Dirichlet-Multinomial and Multinomial Logit models of brand choice". *Marketing Letters, 4 (April), 99-112.*

[8] Goodhardt, G.J., Ehrenberg, A.S.C., et Chatfield, C., (1984), "The Dirichlet: A Comprehensive Model Of Buying Behaviour". *J.R. Statist, Soc, 147, 5, 621-655.*

[9] Hamilton, J.D., (1989), "A new approach to the economic analysis of nontstationary times series and the business cycle". *Econometrica 57, 357-384.*

[10] Henry, O.T., (2009), "Regime switching in the relationship between equity returns and short-term interest rates in the UK". *Journal of Banking & Finance, 33, 405–414.*

[11] Jeuland, A.P., Bass, F.M., et Wright, G.P., (1980), "A Multibrand Stochastic Model Compounding Heterogeneous Erlang Timing and Multinominal Choice Processes". *Operations Research, 28, 255–277.*

[12] Kim, C.J., (1994), "Dynamic linear models with markov-switching". *Journal of Econometrics 60, 1–22.*

[13] Maheu, J., et McCurdy, T., (2000), "Identifying bull and bear markets in stock returns". *Journal of Business and Economic Statistics, 18, 100–112.*

[14] Mandelbrot, B.B., et Taylor, H. W., (1967), "On the Distribution of Stock Price Differences", *Operations Research, 15, 1057-1062.*

[15] Mandelbort, B.B., (1963), "The variation of certain speculative prices", *Journal of Business, 36, 394-419.*

[16] Zhang, F., (2006), "Information Uncertainty and Stock Returns". *Journal of Finance, 61,105–37.*

ESTIMATION BAYESIENNE D'UN MODELE DE VOLATILITE STOCHASTIQUE

Belmokhtar Chouik * et Ouali Anes [o]

Résumé. Les modèles de volatilité stochastique (SV) sont plus difficiles à estimer que les modèles traditionnels de type ARCH. Nous utilisons un algorithme bayesien basé sur des méthodes de Monte Carlo pour estimer les paramètres du modèle (SV). La méthode développée est appliquée à des données simulées.

Abstract. Stochastic volatility models (SV) are more difficult to estimate than the ARCH type models. We use Bayesian algorithm based on the Monte Carlo methods to estimate the parameters of the model (SV). The method developed is applied to simulated data.

Mots-clés : Méthodes Bayesiennes. Econométrie.

1. Introduction

La variance conditionnelle du taux de rendement d'un actif financier, appelée souvent la volatilité, a un impact important sur l'aspect économique et financier. La littérature pour les travaux liés à l'étude de la volatilité est très diversifiée.

Dans les modèles ARCH, introduits par Engle (1982), la volatilité est considérée comme une fonction déterministe. Une extension de ces modèles a été développée dans le but de capturer ce phénomène, nous citons GARCH, EGARCH... L'autre alternative est fournie par les modèles de volatilité stochastique (SV). Cette classe (SV), introduite par Taylor (1982), considère que la volatilité est un processus stochastique latente et offre plus de flexibilité dans la modélisation des données mais présente plus de difficultés dans l'estimation des paramètres.

Le modèle standard de la volatilité stochastique donné par Jacquier E et al. (1994) est donné comme suit :

*** Belmokhtar Chouik,** Ecole Supérieure de Commerce, Alger (b_chouik@esc-alger.com)

[o] **Ouali Anes,** Ecole Nationale supérieure de la statistique et de l'économie appliquée et Laboratoire de Modélisation Stochastique et traitement des données, USTHB, Alger (oualianes@yahoo.fr)

$$y_t = \sqrt{h_t}\,\varepsilon_t \quad (1)$$

$$\log(h_t) = \alpha + \delta \log(h_{t-1}) + \sigma_h \nu_t \quad (2)$$

Les erreurs $\{\varepsilon_t\}$ et $\{\nu_t\}$ sont des bruits blancs gaussiens, $N(0;1)$, qui sont mutuellement indépendants. Le processus $\{h_t\}$ possède un vecteur paramètre $\theta = (\alpha, \delta, \sigma_h)$, où α est le facteur d'échelle de la volatilité, δ représente le paramètre de persistance des chocs et σ_h désigne la volatilité de la log-volatilité.

Plusieurs approches ont été utilisées pour l'inférence statistique des paramètres du modèle (SV), nous citons la méthode des moments abordée par Taylor (1986) et l'approche bayesienne utilisée par Jacquier E et al. (1994), (2004).

Nous reprenons l'approche bayesienne pour une inférence statistique des paramètres du modèle (SV) donné par (1) et (2) sur la base d'une loi a priori gamma-normale pour les paramètres. Les moments a posteriori des volatilités exacts ne pouvant pas être obtenus analytiquement, nous utiliserons les méthodes de Monte Carlo par chaines de Markov (MCMC) basées sur le procédé d'échantillonnage de Gibbs avec la technique d'augmentation des données pour estimer l'espérance et la variance a posteriori des paramètres du modèle à savoir θ et h.

2. Inférence Bayesienne des modèles (SV)

Soient $y = (y_1,..., y_T)'$ et $h = (h_1,..., h_T)'$. Selon l'approche bayesienne, nous sommes intéressés par la distribution conjointe a posteriori $f(\theta, h|y)$. Sous un coût quadratique, une estimation bayésienne d'une fonction φ de (θ, h), soit $\varphi(\theta, h)$, est donnée par

$$E[\varphi(\theta, h|x)] \propto \int \varphi(\theta, h) f(y|\theta, h)\pi(\theta, h) d\theta dh$$

Les intégrales de cette forme ont été source de difficultés dans l'inférence bayésienne, en particulier pour les problèmes de grande dimension. Les méthodes de Monte Carlo nous permettent d'approcher ces intégrales à partir de la suite $(\theta^{(s)}, h^{(s)}), s = 1,..., S$ qui est simulée selon $f(\theta, h|y)$. L'échantillonnage de Gibbs est appliqué aux densités conditionnelles $f(\theta|h, y)$ et $f(h|\theta, y)$.

Nous considérons une distribution a priori gamma-normale pour les paramètres du modèle, soit $\pi(\alpha, \delta, \sigma_h) \propto \sigma_h^{-(\upsilon+2)} \exp(-\frac{(\alpha-\alpha_0)^2 + (\delta-\delta_0)^2 + \omega}{2\sigma_h^2})$, où les paramètres α_0, δ_0 et ω sont connus.

La densité a posteriori $f(\theta|h,y)$ est proportionnelle à $f(y|h,\theta)f(h|\theta)\pi(\theta)$

$$f(\theta|h,y) \propto \left[\prod_{t=1}^{T} f(y_t|h_t,\theta)\right]\left[\prod_{t=1}^{T} f(h_t|h_{t-1},\theta)\right]\pi(\theta)$$

$$f(\theta|h,y) \propto \sigma_h^{-(T+\upsilon+2)}\left[\exp(-\frac{\sum_{t=1}^{T}(\log(h_t)-\alpha-\delta\log(h_{t-1}))^2+(\alpha-\alpha_0)^2+(\delta-\delta_0)^2+\omega}{2\sigma_h^2}\right]$$

Les distributions conditionnelles sont données par,

- $\alpha|\delta,\sigma_h^2,h,y \sim N(\alpha^*,\sigma_\alpha^*)$,

où $\alpha^* = \frac{T\bar{G}+\alpha_0}{T+1}$ et $\sigma_\alpha^* = \frac{\sigma_h}{\sqrt{T+1}}$ avec $G_t = \log(h_t)-\delta\log(h_{t-1})$ et $\bar{G} = \frac{1}{T}\sum_{t=1}^{T} G_t$

- $\delta|\alpha,\sigma_h^2,h,y \sim N(\delta^*,\sigma_\delta^*)$,

où $\delta^* = \frac{\hat{\delta}SC+\delta_0}{SC+1}$, $\sigma_\delta^* = \frac{\sigma_h}{\sqrt{SC+1}}$ et $\hat{\delta} = \frac{\sum_{t=1}^{T} F_t H_{t-1}}{SC}$

avec $SC = \sum_{t=1}^{T}\log^2(h_{t-1})$, $F_t = \log(h_t)-\alpha$ et $H_t = \log(h_t)$

- $\sigma_h^2|\alpha,\delta,h,y \sim i\gamma(\frac{T+\upsilon+1}{2}, \frac{\sum_{t=1}^{T}(\log(h_t)-\alpha-\delta\log(h_{t-1}))^2+(\alpha-\alpha_0)^2+(\delta-\delta_0)^2+\omega}{2})$,

où $i\gamma$ est la distribution gamma inverse.

Les distributions conditionnelles des θ_i. étant usuelles, nous pouvons simuler selon ses lois, ce qui n'est pas le cas pour celle des volatilités. La simulation selon $f(h|\theta,y)$ revient à utiliser l'échantillonnage de Gibbs et simuler selon les distributions conditionnelles $f(h_t|h_{-t},\theta,y_t)$. Cette dernière vérifie,

$$f(h_t|h_{-t},\theta,y_t) = f(h_t|h_{t-1},h_{t+1},\theta,y_t) \propto f(y_t|h_t)f(h_t|h_{t-1},\theta)f(h_{t+1}|h_t,\theta)$$

$$f(h_t|h_{t-1},h_{t+1},\theta,y_t) \propto \left[h_t^{-0.5}\exp(-\frac{y_t^2}{2h_t})\right]\left[h_t^{-1}\exp(-\frac{1}{2\sigma^2}(\log(h_t)-\mu_t)^2)\right] \quad (3)$$

où $\mu_t = \frac{\alpha(1-\delta)+\delta(\log(h_{t+1})+\log(h_{t-1}))}{1+\delta^2}$ et $\sigma^2 = \frac{\sigma_h^2}{1+\delta^2}$

Dans ce travail nous utilisons l'échantillonnage de Gibbs donné par Jacquier et all (1994) dont l'algorithme est,

1- spécifier les valeurs initiales $h^{(0)}$ et $\theta^{(0)}$;

2- pour $s = 1,\ldots,S$

. Simuler $h_t^{(s)}$ selon $f(h_t \mid h_{t-1}^{(s)}, h_{t+1}^{(s-1)}, \theta^{(s-1)}, y)$ pour $t = 1,\ldots,T$

. Simuler $\alpha^{(s)}$ selon $f(\alpha \mid y, h^{(s)}, \delta^{(s-1)}, \sigma_h^{2(s-1)}, y)$

. Simuler $\delta^{(s)}$ selon $f(\delta \mid y, h^{(s)}, \alpha^{(s)}, \sigma_h^{2(s-1)}, y)$

. Simuler $\sigma_h^{2(s)}$ selon $f(\sigma_h^2 \mid y, h^{(s)}, \alpha^{(s)}, \delta^{(s)}, y)$

3- poser $s = s + 1$ et retourner au point 2.

Une simulation directe selon (3), étant difficile, nous proposons une augmentation des données qui consiste à considérer le vecteur aléatoire (h_t, r_t) dont la densité conjointe,

$$f(h_t, r_t) \propto \exp(-r_t)\left[\frac{y_t^2}{2h_t^2}\exp(-\frac{y_t^2}{2h_t})\right] \text{ pour } r_t > \varphi(h_t) \text{ et nulle ailleurs}$$

où, $\varphi(h_t) = \frac{1}{2\sigma^2}(\log(h_t) - \mu_t^*)^2$ et $\mu_t^* = \mu_t + \frac{\sigma^2}{2}$.

Nous notons que les distributions conditionnelles sont usuelles. En effet,

- $\exp(-\frac{y_t^2}{2h_t}) \mid r_t \sim U_{[A_{r_t}, B_{r_t}]}$, où $A_{r_t} = \exp(-\frac{y_t^2}{2a_{r_t}})$ et $B_{r_t} = \exp(-\frac{y_t^2}{2b_{r_t}})$

avec $a_{r_t} = \exp(\mu_t^* - \sigma\sqrt{2r_t})$ et $b_{r_t} = \exp(\mu_t^* + \sigma\sqrt{2r_t})$

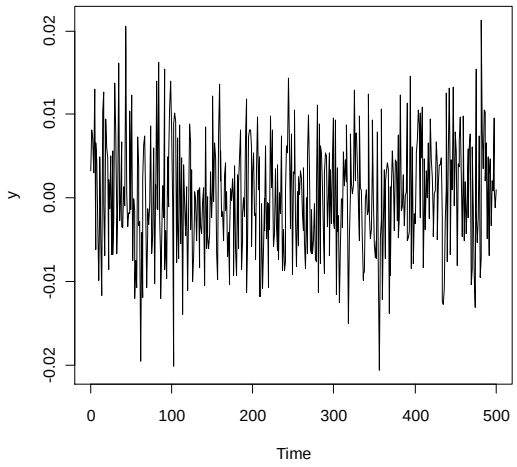
- $r - \varphi(h_t) \mid h_t \sim \xi(1)$.

Ceci nous permet d'appliquer le procédé d'échantillonnage de Gibbs pour générer une chaîne de Markov $\left(h^{(t)}, r^{(t)}\right)$ ayant pour loi stationnaire de densité conjointe $f(h_t, r_t)$; en particulier, $\left(h^{(t)}\right)$ est distribuée selon (3).

3. Application à un modèle SV simulé

Nous générons un processus $\{y_t\}$ selon le modèle donné par (1) et (2), $t = 1,\ldots,500.$ où les vraies valeurs des paramètres α, δ et σ_h sont -0.5, 0.95 et 0.01 respectivement. A l'aide du logiciel R, nous appliquons l'algorithme donné au paragraphe 2 à ces données simulées qui sont représentées graphiquement par la figure 1.

Figure 1: Données simulées selon un modèle SV ($\alpha = -0.5, \delta = 0.95$ et $\sigma_h = 0.01$)



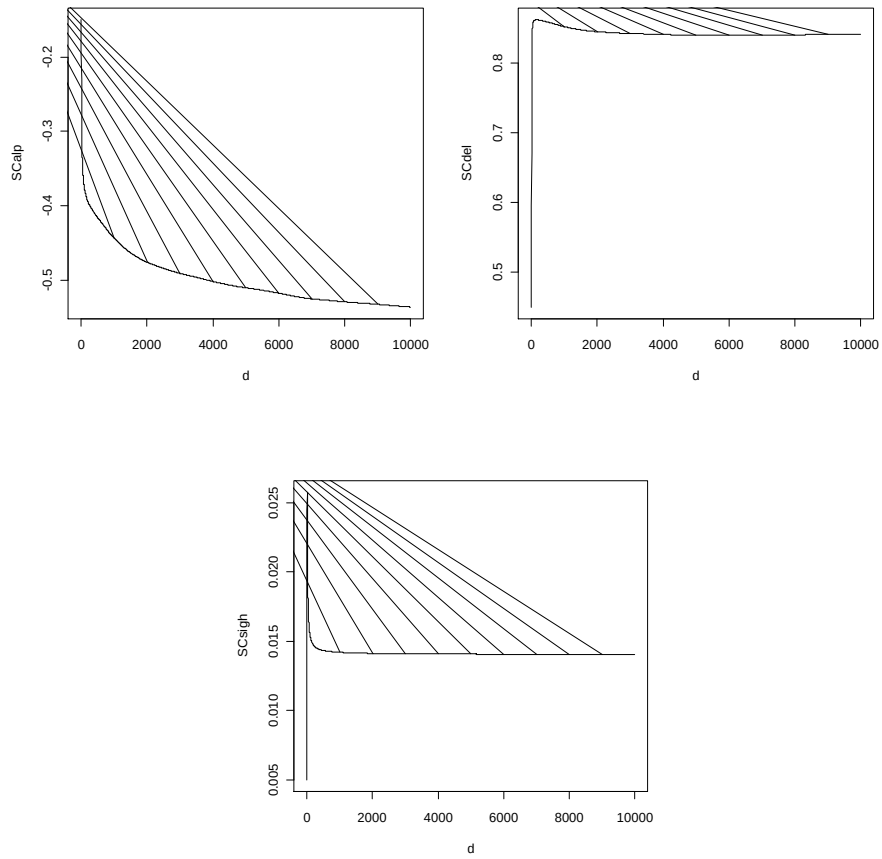
Avec un choix de S=10000, $\alpha_0 = -0.55$, $\delta_0 = 0.9$, $\nu = 10^4$ et $\omega = 1$, les estimations bayesiennes de la moyenne et de l'écart a posteriori, données par le tableau 1, ont été calculées en laissant de côté les 5000 premières simulations. Nous notons que ces estimations sont proches des vraies valeurs.

Tableau 1 : Estimation des moments a posteriori des paramètres

Paramètres	**Vraies valeurs**	**Valeurs estimées**
α	-0.5	-0.561 (0.008)
δ	0.95	0.84 (0.004)
σ_h	0.01	0.014 (0.001)

L'évolution graphique des moyennes cumulées $S_d = d^{-1} \sum_{s=1}^{d} E(\theta_i \left| \theta_{-i}^{(s)} \right.)$ en fonction du nombre d de simulations, où θ_i est une composante de $(\alpha, \delta, \sigma_h)$ et θ_{-i} le reste des paramètres, est représentée par SCalp, SCdel et SCsigh pour les paramètres α, δ et σ_h respectivement. Les convergences des estimateurs peuvent être appréciées par les graphiques de la figure suivante.

Figure 2. Convergence des estimateurs des moyennes a posteriori des paramètres



Bibliographie

[1] Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50 (4):987-1007.

[2] Jacquier E, Polson N. G, Rossi P. E. (1994) Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics.*, 12 (4), 69-87.

[3] Jacquier E, Polson N. G, Rossi P. E. (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics.*, 122, 185-212.

[4] Taylor S. J. (1982). *Financial Returns Modelled by the Product of Two Stochastic Processes: A Study of Daily Sugar Prices*, Volume 1. In Time Series Analysis/ Theory and Practice, o.d, Anderson edition.

[5] Taylor S. J. (1986). *Modelling Financial Times Series,* Wiley.

Résumés des communications

MERCREDI 25 MAI 2011, 10h50

Statistique Publique 2

Statistique et confiance, *Benoit Riandey*

La confiance n'est-elle pas le préalable à de bonnes statistiques ? Confiance des répondants à l'égard de l'institution statistique ou politique, confiance des statisticiens eux-mêmes envers les institutions. La confiance n'est pas moins nécessaire à la bonne réception des travaux statistiques. Cette primauté se voit vite opposer le rôle de l'obligation statistique. Dès l'entre-deux guerres, Alfred Sauvy voyait ses travaux de conjoncture économique entravés par la résistance des entreprises à répondre à ses demandes d'information et, de fait, la loi française de juin 1951 s'intitule Loi relative à l'obligation, la coordination et au secret en matière de statistiques. La place faite à l'obligation est majeure à l'égard des entreprises ; la loi en pose également le principe pour le recensement de population, mais aussi pour les principales enquêtes auprès des ménages. Cette disposition inhabituelle, prise au titre de la la qualité démocratique de la transparence de la société, est souvent critiquée par les collègues anglo-saxons. De fait, en l'absence de confiance, l'obligation de réponse n'induit pas sa sincérité, ce qui appuie notre postulat sur la primauté de la confiance.

La qualité de statistique publique, *Ion Partachi*

Les exigences pour la statistique officielle ont évolué au fil des années, étant exprimées de manière concise et éloquente dans les Principes Fondamentaux de la Statistiques Officielles. Le Code de Bonnes Pratiques des Statistiques Européennes est basé sur 15 principes-clé qui se réfèrent au cadre institutionnel des processus et des produits de la statistique officielle. Si les exigences sur le cadre institutionnel et les processus statistiques viennent d'assurer les conditions nécessaires pour l'élaboration des statistiques de qualité, les rigueurs visant les produits de la statistique officielle stipulent principes et les indicateurs proprement-dits de la "qualité", avec une approche complexe et multidimensionnelle. Toutefois, les utilisateurs sont de plus en plus intéressés par autres aspects de statistiques, qui concernent en grand partie la qualité des services de prestation de l'information statistique par les bénéficiaires : i) la pertinence (le 11-éme principe du Code), ii) la rapidité et la ponctualité (le 13-éme principe), iii) la cohérence et la comparabilité (le 14-éme principe), iv) l'accessibilité et la clarté (le 15-éme principe).

STATISTIQUE ET CONFIANCE

Benoît Riandey

Institut National d'Etudes Démographiques,
133 Bd Davout, 75980 Paris cedex 20 France
riandey@ined.fr

mot clé : statistique publique

La confiance n'est-elle pas le préalable à de bonnes statistiques ? Confiance des répondants à l'égard de l'institution statistique ou politique, confiance des statisticiens eux-mêmes envers les institutions. La confiance n'est pas moins nécessaire à la bonne réception des travaux statistiques.

Cette primauté se voit vite opposer le rôle de l'obligation statistique. Dès l'entre-deux guerres, Alfred Sauvy voyait ses travaux de conjoncture économique entravés par la résistance des entreprises à répondre à ses demandes d'information et, de fait, la loi française de juin 1951 s'intitule *Loi relative à l'obligation, la coordination et au secret en matière de statistiques*. La place faite à l'obligation est majeure à l'égard des entreprises ; la loi en pose également le principe pour le recensement de population, mais aussi pour les principales enquêtes auprès des ménages. Cette disposition inhabituelle, prise au titre de la la qualité démocratique de la transparence de la société, est souvent critiquée par les collègues anglo-saxons. De fait, en l'absence de confiance, l'obligation de réponse n'induit pas sa sincérité, ce qui appuie notre postulat sur la primauté de la confiance.

1- La confiance des répondants

Nous avons évoqué la défiance initiale des entreprises à l'égard des enquêtes statistiques. Le vote de la loi statistique de 1951, l'institution d'un Comité du secret où siègent des représentants patronaux, le fonctionnement statistique durablement sans heurts ont certainement réduit cette défiance. La résistance semble se situer davantage dans la crainte du fardeau statistique, selon l'expression de nos amis québécois.

Qu'en est-il de la confiance des ménages ? Habitués aux enquêtes par le débat public, ils semblent assez confiants à l'égard du secret statistique. Des nuances s'imposent cependant : le taux de refus élevé des indépendants dénote d'une certaine crainte à parler de leurs affaires. Les questions relatives au revenu ou au patrimoine entrainent-elles une dissimulation ? C'est une hantise des statisticiens. Le recensement français ne comprend aucune information sur le revenu et la question générant le plus de crainte est bien surprenante : c'est le nombre de pièces du logement. Tout simplement parce qu'il constitue la base d'imposition des taxes locales.

En fait, le répondant attend surtout d'être traité avec respect. Il s'agit d'abord d'un rapport de confiance inter-personnelle qui implique la loyauté à son égard. Ce principe évident ne s'avère pas si simple dans la pratique ; d'une part, elle peut être mise à mal par des acteurs étrangers à la statistique qui présenteront de façon mensongère leur démarche comme une enquête statistique. De longue date, les instituts de sondage se sont plaints des courtiers d'assurance réputés habitués de cette pratique déloyale. Le démarchage téléphonique a empiré la situation pour les enquêtes téléphoniques. La solution la plus concluante réside dans l'envoi d'une lettre-avis à l'en-tête convaincant pour annoncer l'enquête. Mais surtout les statisticiens d'enquête doivent oeuvrer avec une loyauté intelligente et convaincante : qu'il s'agisse de la durée du questionnaire ou de son contenu, une affirmation maladroite induit la fuite de l'enquêté, la baisse du taux de réponse et bien souvent un biais d'échantillonnage fortement corrélé au sujet de l'enquête. L'enquêté n'étant pas le

demandeur de l'entretien, son accord ne repose pas sur un contrat au sens des juristes mais sur la dynamique d'une confiance qui s'établit au cours de l'entretien.

La confiance accordée à l'enquêteur possède une qualité multiplicatrice dans les enquêtes aréolaires. C'est le cas du recensement, mais également de l'enquête Emploi en France : un questionnaire de taille modeste est ressenti comme respectueux du temps de l'enquêté ; un thème d'enquête non problématique le met à l'aise ; l'enquêté se fait alors l'allié de l'enquêteur auprès des voisins inquiets et l'aide à contacter les personnes peu présentes à leur domicile.

Les sondages électoraux font l'objet d'une assez forte défiance de la part de beaucoup d'enquêtés. Le caractère secret du vote est une garantie de la démocratie. Même après l'énonciation du secret statistique, décliner son vote ou son intention atteste d'une grande confiance de l'enquêté, ou d'une grande indifférence : au cours d'un sondage téléphonique aucunement annoncé par une lettre-avis, qu'est-ce qui témoigne de la réalité du sondage, de la non usurpation d'identité institutionnelle, sinon la conviction que personne ne s'intéresse aux opinions politiques personnelles d'un citoyen tout à fait quelconque ? Cette procédure téléphonique, certes tout à fait rigoureuse de la part des instituts, n'apporte donc pas la garantie de confidentialité de la part de l'appeleur. Il n'est pas surprenant qu'une telle procédure conduise à d'importants biais d'échantillonnage. Ceux-ci sont d'ailleurs difficilement discernables d'une dissimulation des opinions.

D'ailleurs paradoxalement la loi Informatique et Libertés apporte à l'enquêté plutôt confusion que confiance sur les sujets sensibles, tant y apparaît contradictoire une procédure d'accord exprès, c'est àdire écrit, naturelle pour les fichiers aministratifs libres et pour les informations médicales : recevoir une lettre-avis porteuse de la garantie du secret statistique, puis être appelé à signer qu'on assume le risque de répondre à des questions si sensibles était une démarche aberrante. L'INSEE et et les services statistiques ministériels en sont maintenant dispensés par l'article 8 de la loi révisée en août 2004, à l'actif de la directive européenne d'octobre 1995.

La confiance ne favorise pas simplement l'accueil, mais aussi la sincérité des réponses. Les procédures auto-administrées de collecte sont réputées y contribuer fortement (Rogers, 1999). Pour une part, elles relèvent du confort, celui de ne pas énoncer à un enquêteur un acte ou une opinion génante. La saisie par l'enquêté en face à face sur un ordinateur capi-audio d'une information imédiatement cryptée apporte une bonne garantie de confidentialité. La saisie directe sur un site Internet parfait les conditions de confidentialité et favorise la plus grande sincérité... au prix de difficultés d'échantillonnage redoutables. La hantise d'une escroquerie de site Internet demande néanmoins quelques garanties. Or les internautes ne sont pas toujours bons juges du sérieux de l'adresse Internet de l'enquête.

2- La confiance des administrés

Dans les pays développés, de nombreuses statistiques reposent sur des fichiers administratifs. La crise financière, rendant plus aiguë la contrainte budgétaire, ne fait qu'amplifier cette tendance tant en France (SFdS, 2010) qu'à l'étranger (Cloutier, 2010). Mais la qualité de la source est tributaire de la confiance des administrés. Les institutions représentatives de la population jouent à cet égard un rôle primordial. Il peut s'agir d'associations de malades, des droits de l'homme, des parents d'élèves, ou de syndicats professionnels.

Ainsi des associations très prudentes lors du débat relatif au dossier médical personnel s'affirment favorables à un meilleur accès des épidémiologistes aux fichiers médico-administratifs dans de bonnes conditions de sécurité des données (Haut Conseil de santé Publique, 2010).

Au contraire, un conflit concernant l'école maternelle et primaire française illustre bien cette

tension : l'Education nationale s'apprétait à mette en place un fichier normalisé de gestion des élèves pour ces niveaux scolaires comme il en existe déjà pour les collèges et lycées. Un extrait de ces fichiers devait permettre au système statistique d'analyser les trajectoires scolaires des élèves dans les multiples filières du système éducatif. Les craintes des associations de parents d'élèves et des syndicats d'enseignants se sont portées sur les informations relatives aux pays d'origine des élèves et à leur langue maternelle. L'attitude très répressive du gouvernement français à l'égard des immigrés sans papier a fait craindre un usage policier de ces informations en vue d'expulsions de tels parents d'élèves. Un fort mouvement social dans les écoles a permis à ces instances représentatives de faire écarter des fichiers les informations relatives aux origines des élèves. Nul doute cependant de leur intérêt pour la mesure de la réussite scolaire de ces jeunes élèves issus de l'immigration et pour la mesure de l'inadaptation du système éducatif français à ces populations. Le système statistique en est devenu défaillant, limité à des panels scolaires pertinents mais d'une dimension très insuffisante pour explorer l'efficacité des multiples filières à l'égard de cette sous-population et des décrochages dont l'origine peut se situer dans les premières années de scolarisation.

Les enquêtes par sondage devraient échapper à la suspicion que peuvent connaître les fichiers administratifs. Néanmoins en période de climat social tendu, elles peuvent être indûment entrainées dans la bourrasque revendicative ; le doute n'est jamais favorable à une opération de collecte statistique et les mouvements militants, parfois enclins à faire feu de tout bois, ne sont pas toujours informés ou attentifs à distinguer la finalité statistique d'une finalité administrative. La crainte du "fichage" est tout à fait hors de propos pour une enquête par sondage menant à un fichier anonyme bien à l'opposé d'un fichier administratif. C'est une distinction que les statisticiens doivent rappeler avec une pédagogie insistante.

3- La confiance des statisticiens

Les premières personnes à devoir faire preuve de confiance dans les données et les procédures statistiques sont bien les statisticiens eux-mêmes et les autorités publiques. Les statisticiens sont les citoyens les plus à même d'évaluer le degré d'indépendance de la statistique publique. Quand, en 2009 en France, le groupe Lorraine Data publie sa collection d'articles "Le grand trucage", c'est la marque d'un grand désarroi quant à la censure politique de sa hiérarchie, notamment au niveau des directions ministérielles.

Une telle confiance faciliterait le débat public aux moments clés que sont les grands rendez-vous électoraux, et en France particulièrement les élections présidentielles. En 2007, un candidat gouvernemental avançait sur la base de fichiers administratifs que le taux de chômage n'avait jamais été si bas depuis quinze ans, résultat en contradiction avec ceux de l'enquête Emploi. Les statisticiens du ministère du travail attiraient l'attention sur un changement récent de procédure administrative de nature à accélérer les opérations de radiation dans les fichiers. L'institut de statistique a fait preuve d'une prudence excessive et lancé une enquête de contrôle des non réponses à l'enquête qui ne lui a permis qu'après les élections de confirmer la qualité des résultats de l'enquête Emploi (Durier, 2007). Cet épisode regrettable a nui au débat électoral et entrainé une suspicion injustifiée sur les publications de la statistique officielle. Le code des bonnes pratiques de la statistique européenne publié par Eurostat (2005) insiste à juste titre sur le critère de ponctualité qui avait été gravement mis en défaut à cette occasion.

4- La confiance du public

Au même moment, les deux candidats les plus représentatifs se sont prêtés à un dénigrement grossier de la rigueur de l'indice des prix. Certes cet indice méritait une présentation plus pédagogique de la part des statisticiens mais nullement cette démagogie facile en direction d'un

public que l'enseignement scolaire des mathématiques a pu effaroucher. Mettre en cause de façon injustifiée la confiance dans la qualité et l'indépendance de l'institut de statistique fragilise le débat démocratique. Certes l'épisode grec montre bien que ces qualités n'ont pas été suffisantes toujours et partout. Ce comportment hors de propos est d'autant plus grave.

Le public commet bien facilement une confusion entre deux expressions "*Les chiffres mentent*" et "*Bien des menteurs mentent avec des chiffres*". Les sondages d'opinion peuvent pousser dans cette voie un public conscient de la fragilité de telles publications très dépendantes des formulations retenues. Les sondages électoraux génèrent aussi une défiance à l'égard des statistiques car le public est certainement plus conscient des échecs de ces enquêtes que des difficultés méthodologiques qui leur sont très spécifiques et de leurs succès habituels : si les enquêtes relatives aux premiers tours des scrutins se sont souvent avérées fragiles, la précision des estimations aux seconds tours d'élections présidentielles ferait douter de l'utilité même de la notion d'un intervalle de confiance. Ces deux formes d'enquêtes sont à la fois les plus vulgarisées et parmi les plus fagiles. Ce n'est pas un facteur favorable à l'établissement de la confiance envers les statistiques. Probablement l'utilité sociale du recensement, notamment pour les statistiques locales, apporte la meilleure image publique non contestée de la statistique officielle et de son utilité.

Bibliographie

[1] Cloutier M. (2010) Une vision stratégique pour l'utilisation des données administratives à Statistique Canada, *Symposium 2010 de Statistique Canada*, Ottawa.

[2] Data Lorraine (2009) Le grand trucage : comment le gouvernement manipule les statistiques, La Découverte, Paris.

[3] Durier, S., Gonzalez, L., Macario-Rat I. et Thélot H. (2007) Le chômage baisse depuis début 2006. Résultats de l'enquête Emploi, *INSEE première* n° 1146, INSEE, Paris.

[4] Eurostat (2005), Code des bonnes pratiques de la statistique européenne, *http://europa.eu.int/comm/eurostat/quality*.

[5] Haut Conseil de la santé publique (2009) Les systèmes d'information pour la santé publique, www.hcsp.fr/docpdf/avisrapports/hcspr20091111_sisp.pdf, Paris

[6] Rogers S.M, Gribble J.M., Turner Ch. F. et Miller H.G. (1999) Entretiens autoadministrés sur ordinateur et mesure des comportements sensibles, *Population*, n° 2, 231-250,

[7] SFdS groupe Statistique et société (2010) Compte rendu du séminaire Appariements sécurisés du 16 novembre 2010, www.sfds.asso.fr, Paris

La qualité de statistique publique

Prof.univ.,dr. Ion Partachi, Académie d`études économique de la Moldova

Abstract

We live in an ever more complex, increasingly numerous society, in a state of overwhelming changes. So, the biggest challenges which producers of official statistics are now facing are to provide sufficient, quality statistical information for major user groups, but also to support the public with statistical knowledge that would help users to be informed on various issues and to take related decisions based on sound statistical information. To meet this challenge, as well as to increase the efficiency of statistics, the European Statistical System (ESS) pointed a range of vital changes in terms of content and form. Modernising the system of production and dissemination of European statistics is a challenge, but also a great opportunity to improve the quality of Moldovan statistics, including its comparability at the European level to be in line with user's requirements. Moreover, the European integration aspirations of the Republic of Moldova dictate the need for harmonisation of national statistics with the European standards and norms. However, the assimilation of acquis communautaire, which is a constantly moving target, since the European statistical system is in perpetual development, requires a complex approach and strong efforts.

The paper reflects the author's vision on the main activities needed to be fulfilled to better meet the information demands of internal and external users in quality statistics in the European context.

Key words: statistical knowledge, quality, communication, modernisation, harmonisation, innovative approach

JFL: C10, C40, C44

1. Les exigences de la qualité par les produits statistiques

Les exigences pour la statistique officielle ont évolué au fil des années, étant exprimées de manière concise et éloquente dans les Principes Fondamentaux de la Statistiques Officielles, adoptées par la Commission Statistique de l'ONU en avril 1994. Ultérieurement, celles -ci ont été élaborées et précisées dans le Code de Bonnes Pratiques des Statistiques Européennes, adopté le 24 Février 2005 par le Comité de Programme Statistique de l'Union Européenne qui et promulgué à la Recommandation de la Commission Européenne du 25 mai 2005. Le code est basé sur 15 principes-clé qui se réfèrent au cadre institutionnel des processus et des produits de la statistique officielle.

Si les exigences sur le cadre institutionnel et les processus statistiques viennent d'assurer les conditions nécessaires pour l'élaboration des statistiques de qualité, les rigueurs visant les produits de la statistique officielle stipulent principes et les indicateurs proprement-dits de la «qualité», avec une approche complexe et multidimensionnelle.

D'habitude (au sens étroit du mot), la qualité de l'information statistique liées s'associe à la nature objective des données statistiques, c'est-a-dire avec « précisions » de celles-ci (le 12-eme principe du Code). Cet aspect, tient seulement de la qualité des données statistiques, du degré de réflexion adéquate de la situation réelle dans les statistiques.

Toutefois, les utilisateurs sont de plus en plus intéressés par autres aspects de statistiques, qui concernent en grand partie la qualité des services de prestation de l'information statistique par les bénéficiaires: i) *la pertinence* (le 11-éme principe du Code), ii) *la rapidité et la ponctualité* (le 13-éme principe), iii) *la cohérence et la comparabilité* (le 14-éme principe), iv) *l'accessibilité et la clarté* (le 15-éme principe).

2. **Principe de la pertinence**

À cet égard, le défi consiste dans la dimension des informations nécessaire au processus décisionnel dans la sphère économique et sociale en fonction des nécessités ainsi que l'assurance du rapport optimal entre et la rapidité (information à temps) et fiabilité l'apport de la quantité d'informations en correspondance avec les principales priorités de la société pour assurer les procédures décisionnel et la transparence dans l'activité de décision des autorités.

Evidemment, les différentes catégories d'utilisateurs ont des demandes spécifiques. D'une part, le business, les mass-médias, le grand public, d'habitude sont intéressés par les principaux indicateurs statistiques aux courts délais. D'autre part, les autorités et les milieux universitaires exigent des données plus variées et détaillées. Les micros données sont de plus en plus sollicitées pour l'élaboration de des certaines analyses socio-économiques et démographiques de complexes. Evidemment, les institutions de statistique vont déterminer l'opportunité de lancement des recherches statistiques respectives question dans le dialogue avec les partenaires (d'une part les utilisateurs et d'autre part les fournisseurs données), dans la limites des ressources disponibles, en prenant aussi en compte d'autres principes de la statistique officielle, stipulés également dans le Code – celui de l'efficience sur le rapport des coût y compris de la responsabilité de réponse pour les répondants, de la promptitude et de l'assurance de la confidentialité des données individuelles des répondants.

3. Le principe de promptitude et de ponctualité

Selon à ce principe les statistiques doivent être diffusées d'une manière opportune, c'est à dire dans un délai raisonnable et en conformité avec le calendrier d'émission de l'information annoncée à temps.

Evidement, la rapidité de la diffusion de l'information statistique est cruciale pour l'adoption de certaines décisions adéquates au bon moment ; en même temps elle est limitée par la demande d'assurer la précision et la fiabilité des données. L'obtention des informations statistiques avec une grande précision exige des ressources, de temps et de coût plus élevés que les données recueillies avec un degré d'erreur plus haut.

Evidemment, une promptitude meilleure implique des coûts de temps et des ressources financières plus élevés, et a un impact négatif sur la fiabilité des données.

4. Le principe de la cohérence et comparabilité

Ce principe implique: i) le respect de la cohérence interne des données, offrant la possibilité de combiner et d'usage commun et intégrité des données connexes provenant de sources différentes, ii) la comparabilité au niveau intra et inter-états, ainsi que sur l'aspect temporel des données statistiques.

L'assurance la comparabilité statistique au niveau national et international suppose l'utilisation des définitions, des méthodologies, des classifications, des méthodes et des techniques des recherches statistiques et de compilation des indicateurs harmonisés au niveau mondial et stables dans le temps. En outre, certaines divergences d'offre naturelle entre les données provenant de différentes sources ne peuvent pas évitées, celle-ci dicté par le fait que ces indicateurs se réfèrent aux phénomènes aux catégories de différentes données. Par exemple, les données sur le chômage, calculées par l'établissement statistique sur la basée de la méthodologie du Bureau International du Travail, qui se réfère totalement au marché de l'emploi dans le pays seront différentes des données visant le chômage enregistré par l'autorité nationale du travail et de l'emploi.

Ainsi, la nécessitée devient urgent: i) d'avoir une meilleure corrélation avec la production et la diffusion des statistiques dans les institutions responsables de la statistique officielle, ii) de coordonner les activités des institutions au sein du système statistique national, mais aussi des partenaires – des institutions d'état fournisseurs des données administratives.

La statistique élaborée sur les concepts uniformes, harmonisés au niveau national, mais aussi avec la statistique internationale avancée sera, sans aucun doute, un outil plus précieux dans la mesure de l'évolution économique et sociale, mais aussi les prévisions du futur.

5. Le principe d'accessibilité et de clarté

Ce principe exige la diffusion de l'information statistique d'une manière claire et compréhensible, d'une manière transparente, adéquate et convenable pour les utilisateurs des données statistiques.

Fournir des services statistiques de qualité, dans un feedback transparent, prévisible, avec l'information adéquate, en termes raisonnables sur le contenu, les termes et le moyen de diffusion de l'information statistique, ainsi que d'obtenir la réaction inverse des utilisateurs constitue l'élément clé d'assurer la diffusion systématique de la statistique officielle.

L'un des plus précieux de levier pour la base et la prise de décisions sur la gestion du pays, du business, etc., mais aussi un outil extrêmement d'important pour l'évaluation des performances de l'administration publique est surtout l'accès facile et opportun à l'information statistique authentique, complexe, cohérente, comparable au niveau international. Ainsi, l'obligation de l'institution de statistique est de fournir et, en même temps, le droit de chaque membre de la société est d'obtenir l'accès libre à la statistique officielle établie par l'ISO.

L’institution statistique doit traiter tout utilisateur (ou plutôt le consommateur des services publics offertes par l’organisme central de statistique) ainsi que les organismes publics central - les bénéficiaires traditionnels de la statistique officielle. Ca tient donc du devoir ISO d’élaborer les produits statistiques et les méthodes de diffusion de ceux-ci d'une manière qui permettrait d’assurer l’accès égal et facile, conviviales de tous les consommateurs à l’information statistique officielle.

Une culture statistique, promu par l'ISO dans les rangs du grand public sur les éléments statistiques pourraient conduire à la réduction des désaccords du public.

Les tâches de l’institution de statistique officielle en vue d’augmenter la qualité des produits statistiques

Le strict respect des principes de qualité de l'information statistique est essentiel, qui se manifeste dans « la définition de la statistique officielle » comme un bien public élaboré sur la base des méthodes scientifiques, des règles de conduite de façon impartiale et disponible pour toute la société, mais aussi un défi majeur pour les institutions de statistique officielle.

Suite à la tradition historique, conformément à laquelle les organismes de statistiques avaient la fonction de base de répondre aux exigences d’information du gouvernement, l’activité du bureau national de statistique doit mettre l’accent sur la création d’un système développé de diffusion de l’information statistique qui inclurait toutes les catégories d’utilisateurs intéressés.

Bibliographie

1. Brungger H.(2003). Dissemination of official statistics in an environment overload. *Statistics in transition*, November, Vol.6, No 3, pp. 341-351
2. Fellegi, I.P. (2004). Official statistics - pressures and challenges (ISI President’s Invited Lecture, 2003). *International Statistical Review*, 72, 139–155
3. Fisher J.(2007). A statistical system for future generations. *Conference on Modern Statistics for Modern Society,* Parallel Workshop on Innovations in Statistical Systems, Luxembourg
4. Fernandez-Fernandez F. and Museux J.-M.(2010). From knowledge to quality: Contribution of methodology. *European Conference on Quality in Official Statistics*, Session on Quality management frameworks, Helsinki
5. Giovannini, E. (2007). Statistics and Politics in a 'Knowledge Society'. *OECD Statistics Working Paper* No 2007.02, www.oecd.org/statistics
6. Giovannini, E. (2007). Is globalisation a threat to official statistics? *93rd DGINS Conference - The ESS response to globalisation - are we doing enough?* Plenary session on globalisation processes in the field of statistics, Budapest
7. Voineagu V., Dumitrescu I., Stefanescu D.E.(*2009).* European Statistical System. Revista Romana de Statistica, 12, pp. 2-18
8. Decision No 1578/2007/EC of the European Parliament and of the Council of 11 December 2007 on the Community Statistical Programme 2008 to 2012.*OJ L 344*, 28.12.2007, p. 15-43

9. Decision No 1297/2008/EC of the European Parliament and of the Council of 16 December 2008 on a Programme for the Modernisation of European Enterprises and Trade Statistics (MEETS). *OJL* 340, 19.12.2008, p.76-82
10. European Statistics Code of Practice, http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice
11. Statistical work programme of the Commission for 2010, http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/about_eurostat/documents/AWP_2010_final.pdf

Ruptures

Détection de ruptures pour l'estimation de la demande initiale de voyages SNCF, *Abdullah Oueslati*

Pour ajuster au mieux les quotas des différents billets ouverts à la réservation, il faut connaître la "demande initiale" des clients pour chaque type de billet. C'est cette demande en l'absence de quotas, non observée ou partiellement observée, qu'il s'agit d'estimer. L'approche présentée ici utilise un modèle de détection de ruptures poissonnien pour modéliser l'évolution du nombre de réservations. Le nombre de ruptures, leurs dates ainsi que les paramètres de Poisson entre deux dates de rupture successives sont estimés par maximisation de vraisemblance pénalisée à l'aide d'un algorithme de programmation dynamique. Testé sur jeux de données simulés, le modèle conduit à des estimations de qualité. En outre, sur jeux de données réels, la sélection de modèle et les estimations fournissent des informations pertinentes dans le contexte métier.

Test de détection de rupture dans les processus causaux, *Kengne William Charky*

On considère un processus $X = (X_t)_{t\in\mathbb{Z}}$ appartenant à une classe de processus causaux $\mathcal{M}_T(M, f)$. Nous supposons que le modèle dépend d'un paramètre θ_0 et considérons le problème de test de rupture sur le paramètre. La statistique de test est construite à partir des estimateurs de maximum de (quasi)-vraisemblance du paramètre. A un niveau significatif $\alpha \in (0, 1)$, on montre que le seuil asymptotique du test est plus petit que α. Sous une hypothèse alternative d'une rupture dans le modèle, on montre que la statistique de test converge p.s. vers $+\infty$. Quelques résultats de simulation pour les modèles AR(1), ARCH(1) et GARCH(1,1) sont présentés pour montrer la performance de notre procédure et des comparaisons avec d'avec d'autres approches sont faites.

Utilisation du quasi-maximum de vraisemblance pour détecter les ruptures multiples dans des séries chronologiques causales, *Jean-Marc Bardet, William Kengne and Olivier Wintenberger*

This paper is devoted to the off-line multiple change-point detection in a semiparametric framework. The time series is supposed to belong to a large class of models including AR(∞), ARCH(∞), TARCH(∞),... models where the coefficients change at each instant of breaks. The different unknown parameters (number of changes, change dates and parameters of successive models) are estimated using a penalized contrast built on conditional quasi-likelihood. Under Lipshitzian conditions on the model, the consistency of the estimator is proved when the moment order r of the process satisfies $r \geq 2$. If $r \geq 4$, the same convergence rates for the estimators than in the case of independent random variables are obtained. The particular cases of AR(∞), ARCH(∞) and TARCH(∞) show that our method notably improves the existing results.

Joint segmentation of many aCGH profiles using fast group LARS, *Kevin Bleakley and Jean-Philippe Vert*

Nous introduisons un algorithme de groupe LARS qui segment de façon jointe des profils aCGH d'un ensemble de sujets. On pénalise simultanément le quantité de liberté de tous les profils de sauter d'un niveau constant du nombre de copies à un autre, à des endroits génomiques connu sous le nom de points des cassure. Nous montrons que les points de cassure partagé par de nombreux profils différents ont tendance à être trouvé en premier par l'algorithme, même en présence d'importantes quantités de bruit. Des simulations et une mise en oeuvre de l'algorithme sur les profils aCGG de cancer de la vessie sont fournis.

Détection de ruptures pour l'estimation de la demande initiale de voyage Sncf

Abdullah Oueslati

Université Pierre et Marie Curie - Paris VI
Laboratoire de Statistique Théorique
et Appliquée (LSTA), 4 place Jussieu
75013 PARIS

SNCF, Direction Innovation & Recherche
45 rue de Londres
75379 PARIS Cedex 08

Résumé

Pour ajuster au mieux les quotas des différents billets ouverts à la réservation, il faut connaître la "demande initiale" des clients pour chaque type de billet. C'est cette demande en l'absence de quotas, non observée ou partiellement observée, qu'il s'agit d'estimer. L'approche présentée ici utilise un modèle de détection de ruptures poissonnien pour modéliser l'évolution du nombre de réservations. Le nombre de ruptures, leurs dates ainsi que les paramètres de Poisson entre deux dates de rupture successives sont estimés par maximisation de vraisemblance pénalisée à l'aide d'un algorithme de programmation dynamique. Testé sur jeux de données simulés, le modèle conduit à des estimations de qualité. En outre, sur jeux de données réels, la sélection de modèle et les estimations fournissent des informations pertinentes dans le contexte métier.

Mots-clés : Revenue Management, demande dé-contrainte, détection de ruptures, sélection de modèle

Abstract

In order to define the number of various tickets open for booking, prior demand of customers must be reckoned for each type of ticket. This unconstrained demand, which is not observed or partially observed, needs to be estimated. To describe the booking process, we use here a model detecting change-points in the mean of a Poisson process. The number of change-points, the location of the change-points and the Poisson parameters are estimated by means of a dynamic programming algorithm maximizing a penalized likelihood. Tests on simulated and real data sets lead to good estimations and relevant information about the real data sets.

Keywords : Revenue Management, unconstrained demand, change-point detection, model selection

1 Contexte métier et problématique

L'estimation de la "demande initiale" est une problématique inhérente en général au domaine du "Yield-Management" (ou "Revenue Management"), système de gestion des

capacités disponibles et des prix. Né dans le secteur aérien avec la dérèglementation du marché dans les années 1970 aux États-Unis, il s'est étendu à toutes les entreprises dont les produits sont périssables (hôtellerie, spot publicitaire,...). À la SNCF, le "Yield-Management" est un système de contingentement tarifaire permettant d'optimiser le remplissage des trains ainsi que les recettes clients. Or, les données historiques observées, principalement constituées par les nombres de réservations effectuées, ne représentent que les demandes réalisées. Les observations sont influencées par les différentes contraintes imposées aux clients. Nous ne disposons donc pas de la demande première du client. La connaissance de cette *demande initiale* passée pourrait permettre à la fois d'avoir une meilleure lecture de l'impact du "Yield" et d'améliorer les prévisions de la demande future.

Pour traiter cette problématique, on retrouve deux types de méthodes dans les études en Revenue Management. La première consiste à considérer les nombres de réservations mesurés au jour de départ. Ces observations sont censurées dans les cas où la limite de billets ouverts à la réservation a été atteinte. La *demande initiale* est estimée par des algorithmes itératifs de type EM (Weatherford et Pölt, 2002) ou par des modèles de durée (Liu et al., 2002). La seconde approche tient compte de la dynamique des réservations entre l'ouverture à la réservation et le jour de départ du train. Des méthodes de lissage (Crystal et al., 2007) ou des modélisations par processus stochastiques (Lee, 1990) sont développées dans ce contexte.

2 Modélisation par processus de Poisson non homogène et détection de ruptures

La démarche adoptée consiste à estimer la dynamique des réservations par un modèle paramétrique. Les paramètres estimés permettront par la suite de simuler le comportement de réservations et en déduire la *demande initiale*.

Dans le cadre retenu ici, l'évolution temporelle du nombre de réservations au sein d'un train ou d'un ensemble de trains est modélisée par un processus de Poisson non homogène $N(t)$ d'intensité variable $\lambda(t) > 0$. La modélisation consiste à estimer cette fonction λ afin que le processus décrive au mieux la trajectoire de l'évolution des réservations. Dans un premier temps, on suppose que l'intensité est constante par morceaux. Il reste alors à estimer la forme de cette fonction en escaliers à partir d'un échantillon de "montées en charge" observées.

Soit l'échantillon $Y_1, \ldots, Y_n$ représentant une trajectoire d'accroissements de réservations pour les dates $1, 2, \ldots, n$ sur un train et un type de billet donnés. On suppose qu'il existe l dates de "rupture" et que les variables Y_i sont indépendantes et distribuées selon

la même loi de Poisson entre deux dates successives :

$$\begin{aligned} 1 \leq i < k_1 \qquad & Y_i \sim \mathcal{P}(\mu_0) \\ k_1 \leq i < k_2 \qquad & Y_i \sim \mathcal{P}(\mu_1) \\ \vdots & \\ k_{l-1} \leq i < k_l \qquad & Y_i \sim \mathcal{P}(\mu_{l-1}) \\ k_l \leq i \leq n \qquad & Y_i \sim \mathcal{P}(\mu_l) \end{aligned}$$

où $\underline{k_l} = (k_1, \ldots, k_l)$ et $\underline{\mu_l} = (\mu_0, \mu_1, \ldots, \mu_l)$ représentent respectivement les dates de rupture et les paramètres de Poisson sur chaque intervalle entre les dates de rupture.

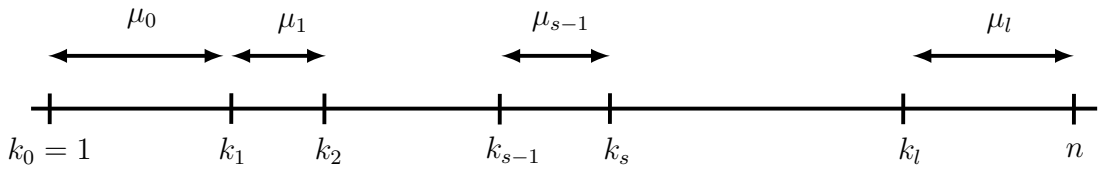


FIGURE 1 – Dates de rupture et paramètres de Poisson

Le but est d'estimer à la fois le nombre de ruptures l, les dates de ruptures $\underline{k_l}$ et les intensités $\underline{\mu_l}$. À l fixé, la log-vraisemblance de l'échantillon comprenant N trajectoires $\underline{Y_1} = (Y_{11}, \ldots, Y_{n1}), \ldots, \underline{Y_N} = (Y_{1N}, \ldots, Y_{nN})$ est :

$$L_l(\underline{k_l}, \underline{\mu_l}) = -N \sum_{s=0}^{l} \mu_s (k_{s+1} - k_s) + \sum_{s=0}^{l} \log(\mu_s) \sum_{i=k_s}^{k_{s+1}-1} \sum_{j=1}^{N} y_{ij}.$$

Pour estimer, à partir d'un jeu de données, le nombre de ruptures l à estimer et les paramètres $\underline{k_l}$ et $\underline{\mu_l}$ correspondants, on maximise la vraisemblance pénalisée de la façon suivante :

$$\widehat{l^c} = \underset{l \in \{1, \ldots, l_{\max}\}}{\operatorname{argmax}} \; L_l - c\, l \log\left(\log\left(nN\right)\right).$$

Le paramètre c optimal, de valeur proche de 2, doit être déterminé par simulations.

3 Mise en œuvre et résultats

La maximisation de la vraisemblance requiert le calcul de la vraisemblance pour chaque combinaison de vecteur $\underline{k_l}$ possible, c'est-à-dire $\binom{n-1}{l}$ calculs de vraisemblance. Ainsi, le

nombre de ruptures l à estimer et la longueur n des trajectoires d'estimation sont limités par les temps de calcul. C'est pourquoi un algorithme de programmation dynamique réduisant grandement le nombre d'opérations ainsi que les temps de calcul a été développé. Le principe de cette méthode est décrit par Bellman (1961). L'algorithme que nous avons développé rejoint celui présenté dans l'article de Jackson et al (2005).

Pour la sélection de modèle, des simulations ont été effectuées pour la calibration du paramètre de pénalisation c. La pénalisation a aussi été effectuée à l'aide d'une heuristique de pente.

Des tests sur données simulées ont fourni des estimations précises sur la loi des variables simulées. En particulier, elles ont confirmé l'efficacité de la procédure de sélection de modèle fixant le nombre de ruptures. La mise en œuvre sur des données réelles a permis en outre de détecter des phénomènes intéressants quant aux comportements de réservation des clients.

Références

[1] L. R. Weatherford et S. Pölt. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1 (3) :234–254, 2002.

[2] P. H. Liu et al. Estimating unconstrained hotel demand based on censored booking data. *Journal of Revenue and Pricing Management*, 1 (2) :121–138, 2002.

[3] C. Crystal et al. A comparison of unconstraining methods to improve revenue management systems. *Production and Operations Management*, 2007.

[4] A. O. Lee. *Airline reservations forecasting : probabilistic and statistical models of the booking process.* PhD thesis, Massachusetts Institute of Technology, 1990.

[5] Y. Gobulev et V. Spokoiny. Exponential bounds for minimum contrast estimators. *Electronic Journal of Statistics*, 3 :712–746, 2009.

[6] R. W. West et R. T. Ogden. Continuous-time estimation of a change-point in a poisson process. *Journal of Statistical Computation and Simulation*, 56 :293–302, 1997.

[7] R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4 (6), 1961.

[8] B. Jackson et al. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12 (2) :105–108, 2005.

[9] A. Aue et al. Estimation of a change-point in the mean function of functionnal data. *Journal of Multivariate Analysis*, 100 :2254–2269, 2009.

[10] V. Spokoiny. Multiscale local change point detection with applications to value-at-risk. *Annals of Statistics*, 37 (3) :1405–1436, 2009.

[11] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51 (4) :1611–1615, 2005.

Test for parameter change in general causal models

Kengne William Charky

SAMM, Université Paris 1, 90 Rue de Tolbiac 75634 Paris Cedex 13, FRANCE

Abstract

We consider a process $X = (X_t)_{t\in\mathbf{Z}}$ belonging to a large class of causal models including AR(∞), ARCH(∞), TARCH(∞),... models. We assume that the model depends on a parameter θ_0 and consider the problem of test for change of the parameter. Test statistic is constructed using quasi-likelihood estimators of the parameter. Given a significance level $\alpha \in (0, 1)$, it is shown that the asymptotic size of the test less than α. Under the local alternative that there is one change, we show that the test statistic converges almost surely to $+\infty$. Some simulation results for AR(1), ARCH(1) and GARCH(1,1) models are reported to show the applicability and the performance of our procedure and comparing to some other approaches.

Keywords : semi-parametric test; Change of parameters; Causal processes; Quasi-maximum likelihood estimator; Weak convergence.

Résumé

On considère un processus $X = (X_t)_{t\in\mathbf{Z}}$ appartenant à une classe de processus causaux $\mathcal{M}_T(M, f)$. Nous supposons que le modèle dépend d'un paramètre θ_0 et considérons le problème de test de rupture sur le paramètre. La statistique de test est construite à partir des estimateurs de maximum de (quasi)-vraisemblance du paramètre. A un niveau significatif $\alpha \in]0, 1[$, on montre que le seuil asymptotique du test est plus petit que α. Sous l'hypothèse alternative d'une rupture dans le modèle, on montre que la statistique de test converge p.s. vers $+\infty$. Quelques résultats de simulation pour les modèles AR(1), ARCH(1) et GARCH(1,1) sont présentés pour montrer la performance de notre procédure et des comparaisons avec d'avec d'autres approches sont faites.

Mots-clés : Test semi-paramétrique ; changement du paramètre ; Processus causal ; Estimateur de maximum de quasi-vraisemblance ; convergence faible.

References

[1] Bardet, J.-M. and Wintenberger, O. (2009) Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. *Ann. Statist. 37*, 2730–2759.

[2] Bardet, J.-M. , Kengne, W. and Wintenberger, O. (2010) Detecting multiple change-points in general causal time series using penalized quasi-likelihood. Preprint available on http://arxiv.org/pdf/1008.0054.

[3] Basseville, M. and Nikiforov, I. (1993) Detection of Abrupt Changes: Theory and Applications. Prentice Hall, Englewood Cliffs, NJ.

[4] Berkes, I., Horváth, L., and Kokoszka, P. (2004) Testing for parameter constancy in GARCH(p; q) models. *Statistics & Probability Letters 70*, 263–273.

[5] Horváth, L., Horváth, Z. and Husková, M. (2008) Ratio tests for change point detection. *Inst. Math. Stat. 1*, 293-304.

[6] Inclan, C., Tiao, G. C. (1994) Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association 89*, 913–923.

[7] Kouamo, O., Moulines, E. and Roueff, F. (2010) Testing for homogeneity of variance in the wavelet domain. *In Dependence in Probability and Statistics*, P. Doukhan, G. Lang, D. Surgailis and G. Teyssiere. Lecture Notes in Statistic 200 , Springer-Verlag, pp. 420–447.

[8] Lee, S. and NA, O. (2005) Test for parameter change in stochastic processes based on conditional least-squares estimator. *J. Multivariate Anal. 93*, 375-393.

[9] Lee, S. , Tokutsu, Y., Maekawa, K. (2004) The CUSUM test for parameter change in regression models with ARCH errors. *Journal of the Japanese Statistical Society 34*, 173–188.

[10] Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika 42*, 523–526.

Utilisation du quasi-maximum de vraisemblance pour détecter les ruptures multiples dans des séries chronologiques causales

Jean-Marc Bardet[1], William Kengne[1] & Olivier Wintenberger[2]

[1] *SAMM, Université Paris 1, 90 Rue de Tolbiac, 75634 Paris Cedex 13, FRANCE*
[2] *CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75016 Paris, FRANCE*

Résumé: Le sujet de cet exposé est celui de la détection de rupture "off-line" dans un cadre semi-parametrique. On considère ainsi des séries chronologiques appartenant à une classe très large de modèles causaux incluant les processus AR(∞), ARCH(∞), TARCH(∞),... On suppose que les coefficients d'un modèle changent à chaque rupture. On notera que comme il peut dépendre d'une infnité de valeurs passées, le modèle est en général non-stationnaire entre chaque instant de rupture. Les différents paramètres inconnus (nombre de ruptures, instant de rupture et paramètres successifs du modèle) sont estimés en utilisant un contraste pénalisé construit à partir d'une quasi-vraisemblance. Sous des conditions lipshitziennes sur le modèle, la convergence de ces estimateurs est montrée dès qu'un moment d'ordre $r \geq 2$ existe. Si $r \geq 4$, les mêmes vitesses de convergence que dans le cas de suites de variables aléatoires indépendantes sont obtenues. Les cas particuliers des AR(∞), ARCH(∞) et TARCH(∞) montrent que notre méthode améliore notablement les résultats existants.

Abstract: This paper is devoted to the off-line multiple change-point detection in a semiparametric framework. The time series is supposed to belong to a large class of models including AR(∞), ARCH(∞), TARCH(∞),... models where the coefficients change at each instant of breaks. The different unknown parameters (number of changes, change dates and parameters of successive models) are estimated using a penalized contrast built on conditional quasi-likelihood. Under Lipshitzian conditions on the model, the consistency of the estimator is proved when the moment order r of the process satisfies $r \geq 2$. If $r \geq 4$, the same convergence rates for the estimators than in the case of independent random variables are obtained. The particular cases of AR(∞), ARCH(∞) and TARCH(∞) show that our method notably improves the existing results.

Mots-Clés: Change detection; Causal processes; ARCH(∞) processes; AR(∞) processes; Quasi-maximum likelihood estimator; Model selection by penalized likelihood.

Bibliographie

[1] Bardet, J.-M. , Kengne, W. and Wintenberger, O. (2010) Detecting multiple change-points in general causal time series using penalized quasi-likelihood. Preprint available on http://arxiv.org/pdf/1008.0054.
[2] Bardet, J.-M. and Wintenberger, O. (2009) Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. *Ann. Statist.*, 37, 2730–2759.
[3] Berkes, I., Horváth, L. and Kokoszka, P. (2003) GARCH processes: structure and estimation. *Bernoulli*, 9, 201–227.
[4] Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2008) Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29, 834–867.
[5] Doukhan, P., and Wintenberger, O. (2008) Weakly dependent chains with infinite memory. *Stochastic Process. Appl.*, 118, 1997–2013.
[6] Francq, C., and Zakoïan, J.-M. (2004) Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10, 605–637.
[7] Kokoszka, P. and Leipus, R. (2000) Change-point estimation in ARCH models. *Bernoulli*, 6, 513–539.
[8] Kounias E. G. and Weng T.-S. (1969) An inequality and almost sure convergence. *Annals of Mathematical Statistics*, 40, 1091–1093.
[9] Lavielle, M. and Ludeña, C. (2000) The multiple change-points problem for the spectral distribution. *Bernoulli*, 6, 845–869.
[10] Lavielle, M. and Moulines, E. (2000) Least squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21, 33–59.

Joint segmentation of many aCGH profiles using fast group LARS

Kevin Bleakley [a] and Jean-Philippe Vert [b,c,d]

[a] INRIA, Saclay, [b]Mines ParisTech, Centre for Computational Biology, Fontainebleau, [c] Institut Curie, Paris and [d] INSERM U900, Paris

kevbleakley@gmail.com, Jean-Philippe.Vert@mines-paristech.fr

Abstract

We introduce a group LARS algorithm that jointly segments aCGH profiles of a set of subjects. It simultaneously penalizes the amount of freedom the set of profiles have to jump from one level of constant copy number to another, at genomic locations known as breakpoints. We show that breakpoints shared by many different profiles tend to be found first by the algorithm, even in the presence of significant amounts of noise. Simulations and an implementation of the algorithm on bladder cancer aCGH profiles are provided.

Résumé

Nous introduisons un algorithme de groupe LARS qui segment de façon jointe des profils aCGH d'un ensemble de sujets. On pénalise simultanément le quantité de liberté de *tous* les profils de sauter d'un niveau constant du nombre de copies à un autre, à des endroits génomiques connu sous le nom de points des cassure. Nous montrons que les points de cassure partagé par de nombreux profils différents ont tendance à être trouvé en premier par l'algorithme, même en présence d'importantes quantités de bruit. Des simulations et une mise en oeuvre de l'algorithme sur les profils aCGG de cancer de la vessie sont fournis.

1 Introduction

Array-based Comparative Genomic Hybridization (aCGH) is a technique that aims to detect copy number variations (CNVs) due to tumor suppressor genes being inactivated by deletion and oncogenes being activated by duplication, on a genomic scale in a single experiment. Many cancers are known to exhibit recurrent CNVs in specific locations in the genome [1, 2, 20, 24, 26].

One challenge related to aCGH profiles is to detect regions of constant copy number, separated by breakpoints, in the presence of significant amounts of noise. Many groups have attempted to answer this question when treating a single profile [5, 8, 9, 13, 15, 16, 23, 25]. Recently, approaches have been suggested for dealing with *multiple* aCGH profiles [3, 12, 14, 19]. Indeed, we often have experimental data from a series of patients who have the same disease status, or several groups of patients, each with a particular disease status. The goal of jointly treating several profiles is to extract breakpoints and copy number variations that are globally representative of the disease status of those patients. Here we present an algorithm that jointly renders a set of n profiles piecewise-constant, with the intended goal of uncovering a set of breakpoints and regions of constant copy number that are pertinent with respect to the whole set of profiles. This algorithm generalizes a Fused Lasso-type algorithm that was used to find breakpoints (and smooth) a single profile in [7].

2 Methods

Suppose we have n aCGH profiles each of length p. The p probe values are calculated at identical locations for each of the n profiles. For each profile, we want to find a piecewise constant representation, where the jumps between constant segments represent *breakpoints*, that is, places where the copy number changes.

2.1 One profile, one chromosome

Our starting point is the following framework for finding breakpoints in a one-dimensional piecewise-constant signal with white noise, as introduced by [7]. Let $Y = (y_1, \ldots, y_p)$ be the observed signal. Consider the following constrained optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \|\beta - Y\|_2^2 \qquad \text{subject to} \qquad \sum_{i=1}^{p} |\beta_i - \beta_{i-1}| < \mu, \tag{1}$$

where μ is a fixed non-negative constant and by convention $\beta_0 = 0$. The constraint $\sum_{i=1}^{p} |\beta_i - \beta_{i-1}| < \mu$ can be seen as a convex relaxation of bounding the number of jumps in β, an idea also implemented in the fused Lasso [23]. When μ is small enough, this constraint causes the solution β to be made up of runs of equally-valued β_i separated by an occasional jump from one constant to another, i.e., a piecewise constant function. For μ large enough the constraint is no longer effective and the solution is merely $\beta = Y$.

This problem is convex, meaning that any standard convex optimization package can solve it for a given μ or a sequence $\{\mu_j\}_{j=1}^{J}$, though this remains computationally intensive. Making the change of variable $u_1 = \beta_1, u_2 = \beta_2 - \beta_1, \ldots, u_p = \beta_p - \beta_{p-1}$, [7] rewrite (1) as

$$\min_{u \in \mathbb{R}^p} \|Au - Y\|_2^2 \qquad \text{subject to} \qquad \sum_{i=1}^{p} |u_p| < \mu,$$

where A is a $p \times p$ matrix whose lower-diagonal and diagonal are 1 and upper-diagonal is 0. This is exactly a Lasso regression problem [22].

2.2 Many profiles, one (or more) chromosomes

Let us now consider a $n \times p$ matrix Y containing n profiles of length p. Our aim is to apply a similar procedure to the one profile case, but jointly to the whole set of profiles. We propose the following constrained optimization problem:

$$\min_{\beta \in \mathbb{R}^{n \times p}} \|\beta - Y\|_2^2 \qquad \text{subject to} \qquad \sum_{i=1}^{p} \|\beta_i - \beta_{i-1}\|_2 < \mu, \tag{2}$$

where μ is a fixed non-negative constant, β_i is the vector of length n containing the value of the i^{th} probe for each of the n individuals and by convention, $\beta_0 = \mathbf{0}$. We now introduce a practical framework for solving such a problem. As for the one profile case [7], we make the change of variable $u_1 = \beta_1, u_2 = \beta_2 - \beta_1, \ldots, u_p = \beta_p - \beta_{p-1}$, where all these objects are now n-dimensional vectors. This gives us the representation:

$$\min_{u \in \mathbb{R}^{np}} \|Au - Y\|_2^2 \qquad \text{subject to} \qquad \sum_{i=1}^{p} \|u_i\|_2 < \mu,$$

where u is the np-dimensional vector of the p vectors u_i of length n stacked on top of each other, by momentary abuse of notation Y is the np-dimensional vector of the columns of Y stacked on top of each other and A is now as follows: for each 1 in the matrix A of the one profile case, replace it with an $n \times n$ identity matrix and for each 0, replace it with an $n \times n$ matrix of zeros. The matrix A thus becomes an $np \times np$ matrix.

In fact, this can be rewritten as a group Lasso [27], i.e.,

$$\min_{u \in \mathbb{R}^{np}} \left\| \sum_{i=1}^{p} A_i u_i - Y \right\|_2^2 \qquad \text{subject to} \qquad \sum_{i=1}^{p} \|u_i\|_2 < \mu, \tag{3}$$

where A_i is the matrix of size $np \times n$ of the columns $n(i-1)+1$ up to ni of A, and u_i the i^{th} column of u. Here, each group i is the set of n variables, one from each profile, found in position i on the genome. Whereas the standard algorithms for solving Lasso and LARS are almost identical, generalizations to

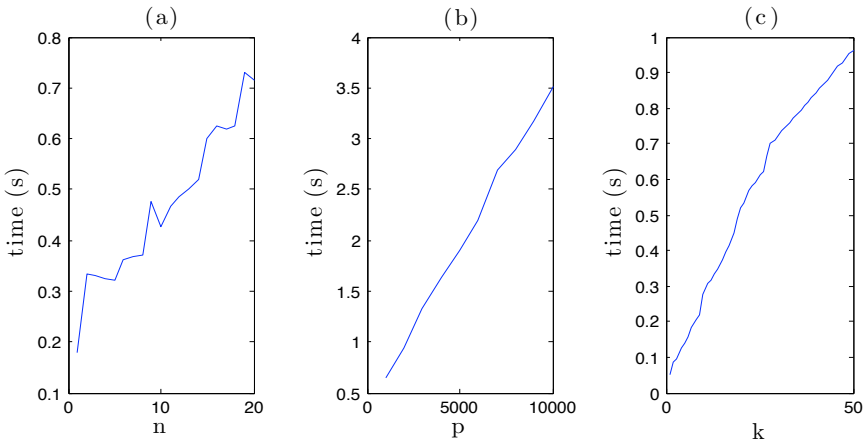


Figure 1: **Speed trials.** (a) CPU time for finding 50 breakpoints when there are 2000 probes and the number of profiles varies from 1 to 20. (b) CPU time when finding 50 breakpoints with the number of profiles fixed at 20 and the number of probes varying from 1000 to 10000 in intervals of 1000. (c) CPU time for 20 profiles and 2000 probes when selecting from 1 to 50 breakpoints.

group Lasso and group LARS are less so. The group LARS algorithm we implemented is therefore not the solution path to (3), but to a similar problem. It explicitly follows the steps given in [27], but avoids the suggested matrix formulation by generalizing the methodology of [7]. A simple modication generalizes the algorithm to multiple chromosomes.

3 Experiments

3.1 Speed trials and performance on simulated data

All trials used Matlab on a 2008 Macbook Pro with 4GB of RAM. Speed trials are shown in Fig.1. Fig. 1(a) indicates linearity in n, and 50 breakpoints were found in 0.72 seconds, an average of 0.014 seconds each. Fig. 1(b) shows linearity in p. Fig. 1(c) shows, for n and p fixed, a near-linear relationship in k, i.e., subsequent breakpoints do not take longer to find than earlier ones. This confirms the theoretical $O(npk)$ complexity.

We performed a series of simulations in order to verify that the algorithm behaved well with respect to our stated goals, namely, recover breakpoints shared by several of a set of profiles. We designed four experiments to move gradually from an artificial to a more realistic setting: (1) All profiles share the same breakpoints; (2) All profiles share the same breakpoint 'regions', though the breakpoints are not all located at exactly the same probe on each profile; (3) All profiles have a subset of a predefined set of breakpoints; (4) All profiles have a subset of a predefined set of breakpoints though the exact location of each breakpoint can vary slightly between profiles . We simulated profiles of length 1000 with 10 breakpoints and varying levels of noise: $\sigma^2 \in \{0.01, 0.1, 0.2, 0.5\}$. Simulation results from experiments 1-2 are shown in Figure 2 (results are analagous in experiments 3 and 4). The main result is that, given enough profiles, the algorithm correctly selected the 10 breakpoint locations/regions for every experimental condition and every noise level.

3.2 Application to bladder tumor CGH profiles

We considered a publicly available aCGH data set of 57 bladder tumor samples [21]. Each aCGH profile gave the relative quantity of DNA for 2215 probes. We removed the probes corresponding to sexual chromosomes, because the sex mismatch between some patients and the reference used made the computation of copy number less reliable, giving us a final list of 2143 probes.

Fig. 3(a) shows the result of superimposing the smoothed versions of the 57 bladder tumor aCGH profiles, when the algorithm has selected 80 ranked common breakpoints. Figs 3(b) and (c) show 2 of the original 57 profiles and their associated smoothed version, where (b) was a profile exhibiting much instability, and (c) only on chromosome 9. We remark that even though (c) was forced to have the

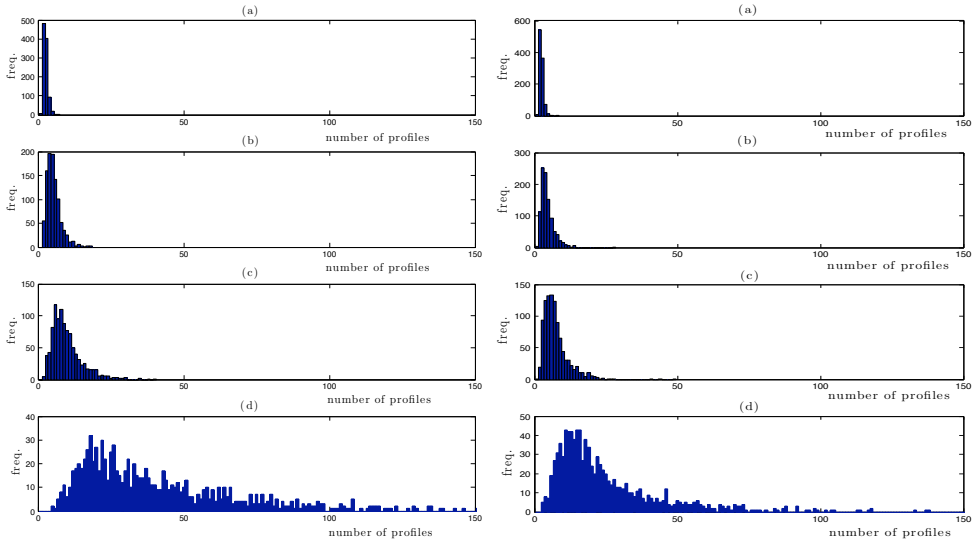


Figure 2: **Simulation conditions 1 (left) and 2 (right).** Left: histograms of the number of profiles required to correctly predict 10 real breakpoints with no mistakes in the presence of white noise. Right: histograms of the number of profiles required to correctly predict all real breakpoints when each profile exhibits a breakpoint in each of 10 tightly defined regions, in the presence of white noise. The noise is $\mathcal{N}(0, \sigma^2)$ with (a) $\sigma^2 = 0.01$, (b) $\sigma^2 = 0.1$, (c) $\sigma^2 = 0.2$ and (d) $\sigma^2 = 0.5$. Each experiment was performed 1000 times.

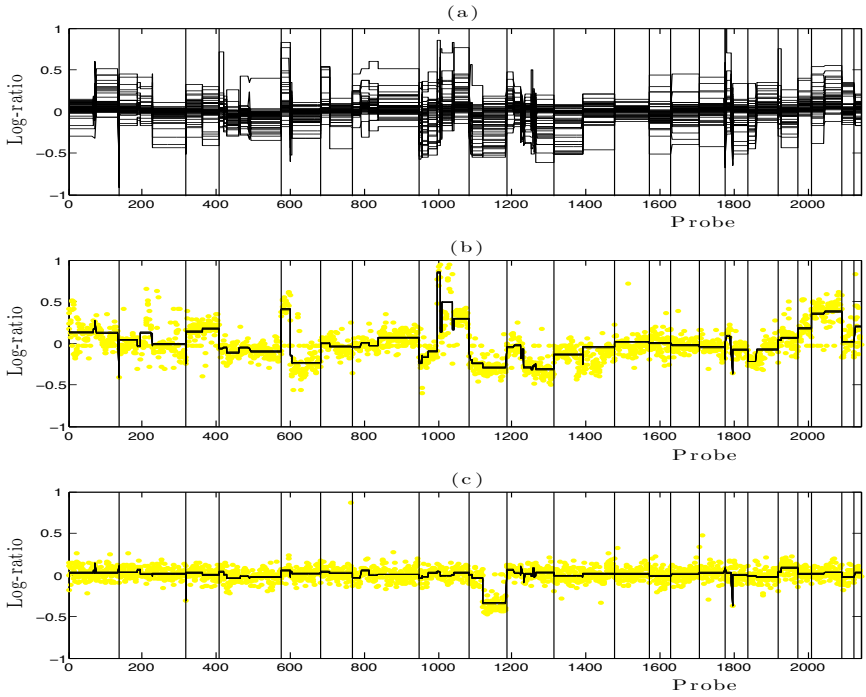


Figure 3: **Graphical representation.** (a) superimposition of the smoothed versions of 57 bladder tumor aCGH profiles [21] with 80 breakpoints. Vertical lines divide chromosomes 1-22. (b) a profile exhibiting many CNVs, and its smoothed version. (c) a profile only showing a deletion on chromosome 9, and its smoothed version. Smoothed profiles are obtained by replacing the set of probe values between consecutive breakpoints with their mean value.

same breakpoints as (b), this does not translate into a poor smoothed version of (c), rather, the forced breakpoints are tiny jumps that can be ignored by biologists.

Fig. 3(a) confirms nearly all of the duplications and deletions associated with bladder cancer found in [1,10,11]: frequent duplication of 8q22-24, 17q21 and 20q is observed, and frequent deletion of 8p22-23, 13q, 17p, 11p and all of chromosome 9. The two known duplications that could not be confirmed here were 12q14-15 and 11q13. Fig. 3(a) suggests other potentially important CNVs, including frequent duplication of 1q, 5p and deletion of 4q and 10q.

4 Discussion

To our knowledge, we have introduced for the first time a way to explicitly code the prior biological information of expecting patients with the same disease to share certain CNVs. Our method forces breakpoints to be located in the same places for *all* profiles. This has the effect of selecting breakpoint locations where many, but not necessarily all, profiles exhibit a breakpoint. This corresponds exactly to one of the underlying biological goals in CNV studies. As shown in Fig. 3(c), it is important to note that a profile forced to have breakpoints where it clearly does not, still ends up with a good quality smoothed representation.

References

[1] E. Blaveri, J. L. Brewer, R. Roydasgupta, J. Fridlyand, S. DeVries, T. Koppie, S. Pejavar, K. Mehta, P. Carroll, J. P. Simko, and F. M. Waldman. Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin Cancer Res*, 11(19 Pt 1):7012–7022, Oct 2005.

[2] N. Bown, M. Lastowska, S. Cotterill, S. O'Neill, C. Ellershaw, P. Roberts, I. Lewis, and A. D. Pearson. 17q gain in neuroblastoma predicts adverse clinical outcome. U.K. cancer cytogenetics group and the U.K. children's cancer study group. *Med. Pediatr. Oncol.*, 36:14–19, 2001.

[3] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Jr Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, 16:1149–1158, 2006.

[4] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani. Least Angle Regression. *Ann. Stat.*, 32(2):407–499, 2003.

[5] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson, and A. Jain. Hidden markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, 90:132–153, 2004.

[6] D. Gershon. DNA microarrays: more than gene expression. *Nature*, 437:1195–1198, 2005.

[7] Z. Harchaoui and C. Lévy-Leduc. Catching change-points with lasso. In *Adv. Neural Inform. Process. Syst. 22*, volume 22, 2008.

[8] Jian Huang, Arief Gusnanto, Kathleen O'Sullivan, Johan Staaf, Ake Borg, and Yudi Pawitan. Robust smooth segmentation approach for array cgh data analysis. *Bioinformatics*, 23(18):2463–2469, Sep 2007.

[9] P. Hupé, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot. array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20:3413–3422, 2004.

[10] A. Kallioniemi, O. P. Kallioniemi, G. Citro, G. Sauter, S. Devries, R. Kerschmann, P. Caroll, and F. Waldman. Identification of gains and losses of DNA sequences in primary bladder cancer by comparative genomic hybridization. *Gene Chromosome Canc*, 12:213–219, 1995.

[11] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258:818–821, 1992.

[12] C. Klijn, H. Holstege, J. de Ridder, X. Liu, M. Reinders, J. Jonkers, and L. Wessels. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, 36(2):e13, 2008.

[13] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004.

[14] F. Picard, É. Lebarbier, E. Budinská, and S. Robin. Joint segmentation of multivariate Gaussian processes using mixed linear models. *Research Report*, 2007.

[15] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27, 2005.

[16] F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation-clustering problem for the analysis of array CGH data. *Biometrics*, 63:758–766, 2007.

[17] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.-L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B.-M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, 1998.

[18] S. Robin and V. T. Stefanov. Simultaneous occurrences of runs in independent Markov chains. *Meth. Comput. Appl. Probab.*, 11(2):267–275, 2008.

[19] C. Rouveirol, N. Stransky, P. Hupé, P. La Rosa, E. Viara, E. Barillot, and F. Radvanyi. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, 22(7):849–856, 2006.

[20] M. Speicher, G. Prescher, S. du Manoir, A. Jauch, B. Horsthemke, N. Bornfeld, R. Becher, and T. Cremer. Chromosomal gains and losses in uveal melanomas detected by comparative genomic hybridization. *Clin. Cancer Res.*, 11:7012–7022, 2005.

[21] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. Gil Diez de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J.-P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi. Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, 38:1386–1396, 2006.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58:267–288, 1996.

[23] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

[24] N. Van Roy, J. Vandesompele, G. Berx, K. Staes, M. Van Gele, E. De Smet, A. De Paepe, G. Laureys, P. van der Drift, R. Versteeg, F. Van Roy, and F. Speleman. Localization of the 17q breakpoint of a constitutional 1;17 translocation in a patient with neuroblastoma within a 25-kb segment located between the accn1 and tlk2 genes and near the distal breakpoints of two microdeletions in neurofibromatosis type 1 patients. *Gene Chromosome Canc*, 35:113–120, 2002.

[25] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. A method for calling gains and losses in array CGH data. *Biostatistics*, 6(1):45–58, Jan 2005.

[26] J. Yao, S. Weremowicz, B. Feng, R. C. Gentleman, J. R. Marks, R. Gelman, C. Brennan, and K. Polyak. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res.*, 66:4065–4078, 2006.

[27] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–68, 2006.

Apprentissage - Classification

Ordonnancement multi-classes : optimalité et premières bornes, *Sylvain Robbiano*

Le but de cet article est de définir les objectifs de l'ordonnancement dans un contexte multi-classes de façon rigoureuse, dans le sillage des résultats récents dans le cadre binaire. On se limite ici au cas où seulement 3 classes sont possible. Sous la condition spécifique de monotonicité des rapport de vraisemblance, des solutions optimales à ce problème d'apprentissage sont décrits dans la situation ordinale, i.e. quand il existe un ordre naturel sur l'ensemble des classes. On introduit la surface ROC et son indicateur synthétique naturel, le volume sous la surface ROC (VUS). Ces critères reflétent les performances d'ordonnancement et seront considérés comme des cibles pour l'optimisation empirique. Enfin, nous dérivons plusieurs bornes dont une borne ponctuel sur la surface ROC et une autre sur le déficit de VUS pour une version plug-in de la fonction de régression.

Classification en maintien Postural, *Christophe Denis*

Notre étude a pour objet de contribuer à l'élaboration d'une notion de style postural, et notamment de classer les patients en terme de maintien postural. Une des conséquences les plus fréquentes du trouble de la posture est la chute qui est une des première causes de mortalité chez les personnes âgées. Nous nous sommes appuyés sur une étude menée sur une population composée de sujets atteints de trouble de la posture et de sujets sains. Les déplacements des centres de gravité des sujets ont été enregistrés sur une plateforme de force au cours de différents protocoles. Nous proposons, dans un premier temps, une procédure de classement des protocoles en terme de maintien postural. Cette procédure est basée sur la méthode du maximum de vraisemblance ciblée. Dans un second temps, nous proposons une procédure de classification des patients s'appuyant sur le classement des protocoles. Pour cela, nous utilisons une méthode adaptative d'agrégation de prédicteurs : le super learning.

Une 2D-réduction de dimension par un estimateur de la distance en probabilité de Patrick Fisher, *Wissal Drira and Faouzi Ghorbel*

Here, we intend to evaluate the performance of the 2D-dimension reduction algorithm obtained by a novel estimator of Patrick Fisher distance based upon the orthogonal series. Such algorithm is compared to Fisher discriminate analysis in the mean of the probability of error which we propose to estimate with the 2D-Kernel probability density function estimate.

Robust Adaboost for data fusion problems, *Afef Ben Brahim and Mohamed Limam*

Dans plusieurs problèmes de classification, les données sont généerées à partir de différentes sources. L'exploitation de toutes les données disponibles est importante pour une prise de décision efficace. La fusion de sources de données hétérogènes pour un même problème est naturel-

lement adaptée aux méthodes d'ensemble puisque différents classifieurs peuvent être construits en utilisant des données obtenues à partir de multiples sources, puis combinées pour avoir un résultat final. Nous proposons des méthodes robustes pour la combinaison des classifieurs, ayant pour objectif de réduire l'effet des classifieurs non fiables dans l'ensemble. Les méthodes proposées montrent une amélioration des résultats de classification.

Ordonnancement multi-classes: Optimalité et premières bornes

Sylvain Robbiano

TELECOM ParisTech, 46 rue Barrault, 75634 Paris cedex 13
sylvain.robbiano@telecom-paristech.fr

Résumé Le but de cet article est de définir les objectifs de l'ordonnancement dans un contexte multi-classes de façon rigoureuse, dans le sillage des résultats récents dans le cadre binaire. On se limite ici au cas où seulement 3 classes sont possible. Sous la condition spécifique de *monotonicité des rapport de vraisemblance*, des solutions optimales à ce problème d'apprentissage sont décrits dans la situation ordinale, *ie* quand il existe un ordre naturel sur l'ensemble des classes. On introduit la surface ROC et son indicateur synthétique naturel, le volume sous la surface ROC (VUS). Ces critères reflétent les performances d'ordonnancement et seront considérés comme des cibles pour l'optimisation empirique. Enfin, nous dérivons plusieurs bornes dont une borne ponctuel sur la surface ROC et une autre sur le déficit de VUS pour une version plug-in de la fonction de régression.

Abstract It is the primary purpose of this paper to set the goals of ranking in a multiple-class context rigorously, following in the footsteps of recent results in the bipartite framework. We limitate this presentation to the case in which the labels can take 3 values. Under specific *likelihood ratio monotonicity* conditions, optimal solutions for this global learning problem are described in the ordinal situation, *i.e.* when there exists a natural order on the set of labels. Criteria reflecting ranking performance such as the ROC surface and its natural summary, the volume under the ROC surface (VUS) are considered as targets for empirical optimization. Finally, we derive several bounds including one on the pointwize ROC deficit and another one on the deficit of VUS for the plug-in regression function.

I INTRODUCTION

Le but de l'ordonnancement est d'apprendre à ranger des observations dans l'ordre des valeurs des étiquettes (inconnues) qui leur sont assignées, basé sur un ensemble d'exemples étiquetés. C'est une question importante dans une grande variété d'applications. En médecine, des règles de décision sont nécessaires dans un cadre multi-classes, les étiquettes correspondant à une gradation ordonné de la maladie (de "pas malade" à "gravement malades ") et des diagnostiques basés sur des statistiques de test sont utilisés pour discriminer entre les états pathologiques, voir Nakas et Yiannoutsos (2004) par exemple. Cette tâche d'apprentissage, à mi-chemin entre la *classification* et l'*estimation de la loi de chaque classe*, présente un défi important pour les statisticiens, précisément en raison

de la nature de l'objet à prédire, un pré-ordre sur l'ensemble des observations éventuelles. Évidemment, il ya plusieurs façons de comparer deux pré-ordres sur un espace qui peut-être continu et la définition des critères d'optimalité ou des mesures de risque dans le cadre de l'ordonnancement n'est pas aussi simple que pour la *classification/ régression* ou l'*estimation de la densité*.

L'angle adopté dans le présent document est le suivant. Son principal objectif est de décrire la situation, en termes de distribution de données, où le graphe ROC ou le volume qu'il définit (généralement appelé le VUS) peut être utilisé pour définir "les règles optimales d'ordonnancement". À cet égard, le cadre à 3 classes est complexe, en contraste avec le cadre binaire où un pré-ordre optimal de l'espace d'entrée existe toujours, ce qui correspond à une courbe ROC qui domine toute autre courbe ROC de manière ponctuelle. Ici, nous montrons qu'une hypothèse, dite *monotonicité des rapport de vraisemblance* sur la collection sous-jacente des lois des classes, garantit l'existence de règles optimales de d'ordonnancement en termes de graphique ROC dans le contexte général.

Le reste du document est structuré comme suit. Dans la section NOTATION, le cadre probabiliste est introduit, avec notations importantes, et la question de l'ordonnacement est formulée d'une manière informelle. Dans la section COURBE ROC ET OPTIMALITÉ, Une hypothèse de *monotonie du rapport de vraisemblance* est présentée, et on montre qu'elle garantit l'existence d'un pré-ordre naturel optimal sur l'espace d'entrée. Dans la section SURFACE ROC. Il est rappelé comment étendre la notion de *graphique* ROC au cadre à 3 classes, et il est établi qu'elle fournit un critère (fonctionnelle) et quantitatives qui permettent d'évaluer la performance d'une règle de d'ordonnancement en vertu de l'hypothèse ci-dessus. Il est également montré que le volume qu'elle définit dans l'espace ROC, appelée *volume sous la surface* ROC (VUS) dans le cadre à 3 classes, peut servir de critère (scalaire) de performance pour l'ordonnancement.

II NOTATIONS

Nous nous plaçons dans le même cadre probabiliste que celui de la *régression ordinale*. Plus précisément, on a un système composé d'une sortie aléatoire, prenant ses valeurs dans un ensemble ordonné discret, $\mathcal{Y} = \{1, 2, 3\}$, et une entrée aléatoire X, à valeurs dans un espace de grande dimension $\mathcal{X}$, modélisant de l'information (que l'on espère pertinente) pour prédire Y. Dans la suite, $F_k(dx)$ désigne la loi conditionnelle de X sachant que $Y = k$, $\mathcal{X}_k$ son support et on pose $p_k = \mathbb{P}\{Y = k\}$ pour $k = 1, 2, 3$. Sans perte de généralités, on suppose que $\mathcal{X}$ coïncide avec $\cup_{k\leq 3}\mathcal{X}_k$. D'autre part, la loi de la couple aléatoire (X, Y) peut être décrit par la loi marginale de X et les probabilités a posteriori: $\eta_k(x) = \mathbb{P}\{Y = k \mid X\}$ avec $x \in \mathcal{X}$ and $1 \leq k \leq 3$ (remarquez que $\sum_{k=1}^{3} \eta_k \equiv 1$). Pour $k \in \{1, 2, 3\}$, on introduit aussi les densités $\Phi_k(X) = dF_k/d\mu(X)$, ainsi que les rapports de vraisemblance (éventuellement infini) $\Phi_{k,l}(X) = \frac{dF_k}{dF_l}(X) = \frac{\Phi_k}{\Phi_l}(X)$, avec $1 \leq k,\ l \leq K$ et la convention $u/0 = \infty$ pour tout $u \in]0, \infty[$ et $0/0 = 0$.

Ces quantités sont reliées les unes aux autres par les équations: $\mu(dx) = \sum_{k=1}^{3} p_k \cdot F_k(dx)$ et, pour $1 \leq k,\ l \leq 3$, $\eta_k(X) = p_k \cdot \Phi_k(X)$ et $\Phi_{k,l}(X) = (p_l/p_k) \cdot \eta_k(X)/\eta_l(X)$.

L'espérance conditionnelle de la sortie v.a. Y sachant X est notée :
$\eta(X) = \mathbb{E}[Y \mid X] = \sum_{k=1}^{3} k \cdot \eta_k(X)$.

Dans la suite, on notera $\mathcal{S}$ l'ensemble des fonctions boréliennes $s : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. Ses éléments seront appelés *fonctions de scoring*. Notez que la valeur $+\infty$ est autorisée, de sorte que les rapports de vraisemblance peuvent être considérés comme des fonctions de scoring. Enfin, Δ désigne la différence symétrique entre deux ensembles, $\mathbb{I}\{\mathcal{E}\}$ la fonction indicatrice de tout événement $\mathcal{E}$.

III COURBE ROC ET OPTIMALITÉ

Ici on rappelle la définition de la courbe ROC et les concepts associés pour en déduire l'optimalité dans le cas multi-classes.

Définition 1 (ROC CURVE) *Soient $F_1(dx)$ et $F_2(dx)$ deux distributions de probabilité sur $\mathcal{X}$. La courbe* ROC *d'une fonction de scoring $s : \mathcal{X} \to]-\infty, +\infty]$ par rapport à la paire (F_1, F_2) est la courbe paramétrée*

$$t \in \mathbb{R} \mapsto \left(\mathbb{P}\left\{s(X) > t \mid Y = 1\right\}, \mathbb{P}\left\{s(X) > t \mid Y = 2\right\}\right).$$

La ROC courbe peut être considérée comme le graphe d'une certaine fonction croissante *càd-làg* : $\alpha \in [0,1] \mapsto \mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{s}, \alpha)$, définie par

$$\mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{s}, \alpha) = 1 - \mathrm{F_{s,2}} \circ \mathrm{F_{s,1}^{-1}}(1-\alpha)$$

aux points α tels que $F_{s,1} \circ F_{s,1}^{-1}(1-\alpha) = 1-\alpha$, où $W^{-1}(u) = \inf\{t \in]-\infty, +\infty] : W(t) \geq u\}$, $u \in [0,1]$, est l'inverse généralisée de toute fonction de répartition $W(t)$ sur $\mathbb{R} \cup \{+\infty\}$. Nous nous référons à l'annexe A de Clémençon et Vayatis (2009) dans pour une liste de propriétés des courbes ROC.

Equipé de ce concept, notez que $F_{s,1} \leq_{sto} F_{s,2}$ signifie que la courbe est au-dessus de la première diagonale du carré unité et elle coïncide avec celle-ci lorsque $F_1 = F_2$, quelque soit $s(x)$. L'espace ROC induit un ordre partiel sur l'ensemble des fonctions de scoring: par rapport à la paire (F_1, F_2), une fonction de scoring $s(x)$ est moins performante qu'une autre $s'(x)$ quand: $\forall \alpha \in [0,1]$, $\mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{s}, \alpha) \leq \mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{s}', \alpha)$. En ce qui concerne cette façon d'évaluer la performance d'ordonnancement dans la situation binaire, l'ensemble $\mathcal{S}^*_{F_1,F_2}$ des fonctions de scoring optimales est l'ensemble des fonctions $s \in \mathcal{S}$ telles que: $\forall (x, x') \in \mathcal{X}^2 : \quad \Phi_{F_2,F_1}(x) < \Phi_{F_2,F_1}(x') \Rightarrow s(x) < s(x')$. Elle peut être établie par des arguments de Neyman-Pearson, nous nous référons à la proposition 4 dans Clémençon et Vayatis (2009) pour plus de détails. Pour revenir au cadre multi-classes, conformément à l'objectif d'ordonnancement à 3-classes décrit ci-dessus d'une manière informelle, les fonctions optimales de scoring sont naturellement définies comme celles appartenant à l'ensemble $\mathcal{S}^* \stackrel{def}{=} \bigcap_{1 \leq k < l \leq 3} \mathcal{S}^*_{k,l}$, où l'on pose $\mathcal{S}^*_{k,l} = \mathcal{S}^*_{F_k,F_l}$ pour simplifier la notation (remarquer que $\mathcal{S}^* = \bigcap_{1 \leq k < 3} \mathcal{S}^*_{k,k+1}$). On peut construire des exemples tels que l'ensemble $\mathcal{S}^*$ peut être vide. L'hypothèse suivante, qui stipule que tous les rapports $\Phi_{k,l}$, $1 \leq l <$

$k \leq K$, croissent ou décroissent ensemble, peut être facilement considérée comme une condition nécessaire et suffisante pour que $\mathcal{S}^*$ soit non vide.

Hypothèse 1 *Pour tout* $(k,l) \in \{1,2\}^2$*, pour tout* $(x,x') \in \mathcal{X}^2$*, on a:*

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow \Phi_{l+1,l}(x) \leq \Phi_{l+1,l}(x').$$

Le résultat suivant justifie la nature même de cette hypothèse.

Théorème 1 *L'ensemble* $\mathcal{S}^*$ *est vide non si et seulement si l'hypothèse 1 est satisfaite. Dans un tel cas, nous avons nécessairement*

$$\mathcal{X}_{k'} \cap \mathcal{X}_{l'} \subset \mathcal{X}_k \cap \mathcal{X}_l \text{ for any } 1 \leq k' \leq k < l \leq l' \leq 3.$$

En outre, l'ensemble $\mathcal{S}^*$ *est l'ensemble des fonctions de scoring* $s \in \mathcal{S}$ *telles que,*

$$\forall (x,x') \in \mathcal{X}^2 : \ \exists k \in \{1,2\} \text{ tel que } \Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow s(x) < s(x').$$

Le résultat suivant montre que l'hypothèse 1 revient à supposer que la famille de densités $\{\Phi_k(x) : \ 1 \leq k \leq K\}$ a ses *rapports de vraisemblance monotones.*

Proposition 1 *L'hypothèse 1 est satisfaite si et seulement si il existe une fonction borélienne à valeurs réelles* $s^*(x)$ *telle que pour tout* $k < l$ *in* $\{1,2,3\}$*, le rapport* $\Phi_{l,k}(x)$ *est une fonction croissante de* $s^*(x)$*. Dans ce cas, la fonction scoring* $s^*(x)$ *appartient à l'ensemble* $\mathcal{S}^*$*.*

Le théorème suivant explicite une fonction de scoring optimale dans le cas où il en existe.

Théorème 2 *Si l'hypothèse 1 est satisfaite, la fonction de régression* $\eta(x)$ *appartient alors à l'ensemble des fonctions optimales de scoring* $\mathcal{S}^*$*.*

Bien que attendu, ce résultat est cruciale et car il donne la fonction cible nécessaire pour les approches de types plug-in pour résoudre la tache d'ordonnancement multi-classe.

IV SURFACE ROC

Suivant les traces de Scurfield (1996) dans les situations avec plus de deux classes, le graphique ROC d'une fonction de scoring $s(x)$ devient l'ensemble des points

$$M(\mathbf{t}) = \left(F_{s,1}(t_1) - F_{s,1}(t_0), F_{s,2}(t_2) - F_{s,2}(t_1), F_{s,3}(t_3) - F_{s,3}(t_2)\right), \tag{1}$$

où $-\infty = t_0 < t_1 \leq t_2 < t_3 = \infty$. Remarquez que $F_{s,3}(t_3) = 1$ et $F_{s,1}(t_0) = 0$ et que les coordonnées du point(1) coïncide avec la diagonale de la *matrice de confusion* de la règle de classification définie par seuillage $s(x)$ à l'échelle t_k, $1 \leq k < 3$, $C_{s,\mathbf{t}}(x) = \sum_{k=1}^{3} k \cdot \mathbb{I}\{t_{k-1} < s(x) \leq t_k\}$. Nous avons en effet $\mathbb{P}\left\{C_{s,\mathbf{t}}(X) = k \mid Y = k\right\} = F_{s,k}(t_k) - F_{s,k}(t_{k-1})$ pour tout k dans $\{1,2,3\}$.

La surface ROC contient clairement toutes les informations portées par les trois courbes $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s},.)$, $\mathrm{ROC}_{\mathrm{F}_2,\mathrm{F}_3}(\mathrm{s},.)$ et $\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},.)$. En particulier, l'intersection de la surface avec chacune des trois faces de la orthant positive coïncide avec l'image d'une de ces courbes par une simple transformation. Le graphe ROC est alors une *surface paramétrique*, qui coïncide avec le graphique

$$\left\{(\alpha, \mathrm{ROC}(\mathrm{s},\alpha,\gamma),\gamma) :\ (\alpha,\gamma)\in[0,1]^2 \text{ such that } \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)\right\}$$

d'une application $\mathrm{ROC}(\mathrm{s},.,.)$ définie sur l'ensemble $\mathcal{I}_s \stackrel{def}{=} \{(\alpha,\gamma)\in[0,1]^2 :\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)\}$ et telle que, à tout point (α,γ) pour lequel $\gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s},1-\alpha)$, $F_{s,1}\circ F_{1,s}^{-1}(\alpha)=\alpha$ et $F_{s,3}\circ F_{3,s}^{-1}(\gamma)=\gamma$, on a:

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = \mathrm{F}_{2,\mathrm{s}}\circ \mathrm{F}_{3,\mathrm{s}}^{-1}(1-\gamma) - \mathrm{F}_{2,\mathrm{s}}\circ \mathrm{F}_{1,\mathrm{s}}^{-1}(\alpha). \tag{2}$$

Dans le théorème suivant, on justifie l'approche ROC de l'ordonnancement car les fonctions optimales de scoring maximise en tout point le graphique ROC.

Théorème 3 *Supposons que l'hypothèse 1 est respectée et posons* $\mathrm{ROC}^*(.,.) = \mathrm{ROC}(\mathrm{s}^*,.,.)$ *pour* $s^*\in\mathcal{S}^*$. *On a, pour toute fonction de scoring* $s\in\mathcal{S}$ *et pour tout* $(\alpha,\gamma)\in[0,1]^2$,

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) \leq \mathrm{ROC}^*(\alpha,\gamma).$$

En outre, si nous fixons, pour tout $\alpha\in[0,1]$, $k\in\{1,\ 2,\ 3\}$ *et* $s\in\mathcal{S}$,

$$R_{s,\alpha}^{(i)} = \{x\in\mathcal{X} | s(x) > Q^{(i)}(s,\alpha)\},$$

où $Q^{(i)}(s,\alpha)$ *désigne le quantile d'ordre* α *de la loi conditionnelle de* $s(X)$ *sachant* $Y=i$ *et supposons que* $\eta(X)$ *est une variable aléatoire continue, on a:* $\forall(\alpha,\gamma)\in[0,1]^2$,

$$\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma) \leq \mathbb{I}\{(\alpha,\gamma) :\ \gamma \leq \mathrm{ROC}^*_{\mathrm{F}_1,\mathrm{F}_3}(1-\alpha)\}\cdot\left(\Theta_1(\mathrm{s},\alpha)+\Theta_2(\mathrm{s},\gamma)\right),$$

où

$$\begin{aligned}
\Theta_1(s,\alpha) &= \frac{\mathbb{I}\{\alpha\neq 0\}}{p_2 Q^{(1)}(\eta_1,\alpha)}\mathbb{E}\left[\left|\eta_1(x) - Q^{(1)}(\eta_1,\alpha)\right|\cdot\mathbb{I}\{R^{(1)}_{s^*,\alpha}\Delta R^{(1)}_{s,\alpha}\}\right],\\
\Theta_2(s,\gamma) &= \frac{\mathbb{I}\{\gamma\neq 1\}}{p_2 Q^{(3)}(\eta_3,1-\gamma)}\mathbb{E}\left[\left|\eta_3(X) - Q^{(3)}(\eta_3,1-\gamma)\right|\cdot\mathbb{I}\{R^{(3)}_{s^*,1-\gamma}\Delta R^{(3)}_{s,1-\gamma}\}\right].
\end{aligned}$$

Bien que pas très lisible dans son énoncé, il est important de noter que la borne ponctuelle ci-dessus dépend explicitement de la différence symétrique des ensembles de niveaux des deux fonctions de scoring. Ce théorème nous permet donc de penser qu'un algorithme qui retrouve les ensembles de niveaux d'une fonction optimale de scoring va être pertinent pour la tache d'ordonnancement. Dans le cas binaire, cette approche a été utilisé pour faire l'algorithme TreeRank(http://treerank.sourceforge.net/).

Proposition 2 (LE CRITÈRE VUS) *Soit $s(x)$ une fonction de scoring. Le volume sous sa surface* ROC *est* $\mathrm{VUS(s)} = \int\int \mathrm{ROC(s}, \alpha, \gamma)\mathrm{d}\alpha\mathrm{d}\gamma$. *Sous l'hypothèse 1, on a:* $\forall s \in \mathcal{S}$, $\mathrm{VUS(s)} \leq \mathrm{VUS}^*$, *avec* $\mathrm{VUS}^* = \mathrm{VUS(s}^*)$ *for* $s^* \in \mathcal{S}^*$.

Proposition 3 (SCURFIELD (1996)) *Pour toute fonction de scoring $s \in \mathcal{S}$ telle que $\mathbb{P}(s(X) = s(X')) = 0$, on a:*

$$\mathrm{VUS(s)} = \mathbb{P}\left\{\mathrm{s(X_1)} < \mathrm{s(X_2)} < \mathrm{s(X_3)} | \mathrm{Y_1} = 1, \mathrm{Y_2} = 2, \mathrm{Y_3} = 3\right\}$$

où (X_1, Y_1), (X_2, Y_2) et (X_3, Y_3) sont des copies indépendantes de la paire aléatoire (X, Y).

Cette propositon est très intéressante car elle donne une forme probabiliste au critère VUS. On voit ainsi que la version empirique du VUS s'exprime comme une U-statistique et que des algorithmes du type maximisation du VUS empirique seront performant. Le théoreme suivant permet de majorer le déficit de VUS par la somme de deux déficits d'AUC.

Théorème 4 (DÉFICIT DE VUS) *Supposons que l'hypothèse 1 est vraie. Alors, pour toute fonction de scoring $s \in \mathcal{S}$, on a*

$$\mathrm{VUS}^* - \mathrm{VUS(s)} \;\leq\; \left(\mathrm{AUC}^*_{\mathrm{F_1,F_2}} - \mathrm{AUC}_{\mathrm{F_1,F_2}}(\mathrm{s})\right) + \left(\mathrm{AUC}^*_{\mathrm{F_2,F_3}} - \mathrm{AUC}_{\mathrm{F_2,F_3}}(\mathrm{s})\right).$$

Ce théorème permet d'imaginer que des algorithmes de type aggrégation de règles de scoring binaire (ici 1 contre 2 et 2 contre 3) font être pertinent pour la tache d'ordonnacement.

Finalement, cette dernière proposition dont la démonstration est basée sur le théorème précèdent, donne une première justification théorique à l'approche plug-in.

Proposition 4 (DÉFICIT DE VUS (BIS)) *Supposons que l'hypothèse 1 est satisfaite. Soit $\widehat{\eta}$ une fonction qui approxime η. Supposons que les variables aléatoires $\eta(X)$ et $\widehat{\eta}(X)$ sont continues. On a:* $\forall s \in \mathcal{S}$,

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \frac{\mathrm{p_1} + \mathrm{p_3}}{\mathrm{p_1 p_2 p_3}} \cdot \mathbb{E}\left[|\eta(\mathrm{X}) - \widehat{\eta}(\mathrm{X})|\right]$$

Bibliographie

[1] Clémençon, S. et Vayatis, N. (2009) *Tree-based ranking methods*, IEEE Transactions on Information Theory, 55,4316-4336.

[2] Nakas, C. et Yiannoutsos, C. (2004) *Ordered multiple-class ROC analysis with continuous measurements*, Statistics in Medicine, 23, 3437-3449.

[3] Scurfield, B.K. (1996) *Multiple-event forced-choice tasks in the theory of signal detectability*, Journal of Mathematical Psychology, 40, 253-269.

Classification en maintien postural

Christophe Denis

MAP5-UMR CNRS 8145, Université Paris Descartes

Résumé

Notre étude a pour objet de contribuer à l'élaboration d'une notion de "style postural", et notamment de classer les patients en terme de maintien postural. Une des conséquences les plus fréquentes du trouble de la posture est la chute qui est une des première causes de mortalité chez les personnes âgées. Nous nous sommes appuyés sur une étude menée sur une population composée de sujets atteints de trouble de la posture et de sujets sains. Les déplacements des centres de gravité des sujets ont été enregistrés sur une plateforme de force au cours de différents protocoles. Nous proposons, dans un premier temps, une procédure de classement des protocoles en terme de maintien postural. Cette procédure est basée sur la méthode du maximum de vraisemblance ciblée. Dans un second temps, nous proposons une procédure de classification des patients s'appuyant sur le classement des protocoles. Pour cela, nous utilisons une méthode adaptative d'agrégation de prédicteurs : le super learning.

Abstract

Our study contributes to the search for a notion of "postural style", focusing on the issue of classifying patients in terms of postural maintainance. The principal consequences of deficit in postural maintainance is fall, which is particularly bad in seniors. Our approach is based on a study on a group of patients. This group is composed by patients who show deficit in postural maintainance and normal patients. Each patient undergoes several protocols. A forceplatform records over time the center-of-pressure of each foot. In a first time, we propose to rank the protocols in terms of postural maintainance. Our ranking of the protocols relies on the targeted maximum likelihood methodology. Then, we propose to classify the patients. Our classification procedure is based on the ranking of the protocols. Classifiers constructions relies on the super learning method of classifiers aggregation.

Mots clés: maintien de la posture, maximum de vraisemblance ciblé, classification, super learning, leave-one-out

1 Introduction

Le maintien postural est le fruit du traitement dynamique de trois types d'informations encodés par les systèmes visuel, vestibulaire (situé dans l'oreille interne) et proprioceptif (composés par les récepteurs sensoriels situés au voisinage des os, des articulations et des muscles et sensibles aux stimulations produites par les mouvements du corps). Les informations proprioceptives sont relatives à la perception de la position, de l'emplacement, de l'orientation et du mouvement des différents membres du corps. Les informations vestibulaires sont elles relatives au sens de l'équilibre.

Naturellement chaque individu a développé, selon son expérience sensorimotrice, ses propres préférences sensorielles. Souvent, un seul type d'information sensorielle est sollicité de façon prédominante. La préférence visuelle est sans doute la plus fréquente. Si une telle sélection systématique d'un unique mode perceptif permet néanmoins de se déplacer efficacement dans un environnement familier, il est clair qu'elle est peu adaptée à la gestion des situations nouvelles ou inattendues. Ce mode de fonctionnement est donc plus susceptible de conduire à une chute, qui peut s'avérer particulièrement dramatique pour les personnes âgées.

Notre étude a pour objet de contribuer à l'élaboration d'une notion de "style postural", et notamment de déterminer quelles sont les informations sensorielles qui prédominent chez les patients atteints de troubles du maintien postural. À long terme, l'espoir est de permettre la mise en place de protocoles de physiothérapie adaptés au trouble de la posture chez chaque patient.

2 Description des données

Les données sur lesquelles notre étude repose ont été collectées par le CESEM (UMR CNRS 8194, Université Paris Descartes). Un total de 54 individus a été suivi lors de l'étude; 22 d'entre eux étaient hémiplégiques et 32 d'entre eux ne présentaient aucun trouble de la posture. Pour chacun des patients un ensemble de covariables (âge, genre, latéralité, taille, poids) a été recueilli. Chacun des individus a été soumis à quatre protocoles différents dont nous donnons un descriptif en Table 1. Pour chaque protocole, le centre de pression de chaque pied a été enregistré au cours du temps sur une plateforme de force. Le résultat d'un protocole est donc composé d'une trajectoire $(X_t)_{t\in T} = (L_t, R_t)_{t\in T}$ où $L_t = (L_t^1, L_t^2) \in \mathbb{R}^2$ (respectivement, $R_t = (R_t^1, R_t^2)$) donne la position du centre de pression du pied gauche (droit) sur la plateforme de force au temps t, pour tout $t \in T = \{k\delta : 1 \leq k \leq 2800\}$ où $\delta = 0.025$ secondes (un protocole dure 70 secondes). La Figure 1 propose une description visuelle d'une trajectoire issue du protocole 3 (voir Table 1) associée à un sujet sain.

La Figure 1 confirme l'intuition que la structure d'une trajectoire générique $(X_t)_{t\in T}$ est compliquée. Nous proposons d'utiliser comme Chambaz, Bonan et Vidal (2009), pour la suite de l'étude, des mesures résumées de $(X_t)_{t\in T}$. Nous procédons à un premier

protocole	1ère phase (0→15s)	2ème phase (15→50s)	3ème phase (50→70s)
1	pas de perturbation	yeux fermés	pas de perturbation
2		stimulation musculaire	
3		yeux fermés stimulation musculaire	
4		stimulation optocinétique	

Table 1: Caractéristiques des quatre protocoles. Un protocole est divisé en trois phases: une première phase sans perturbation de la posture, une deuxième phase avec perturbation et une dernière phase sans perturbation. Différents types de perturbations sont envisagés. Elles peuvent être de nature visuelle (yeux fermés), proprioceptive (stimulation musculaire) ou encore vestibulaire (stimulation optocinétique).

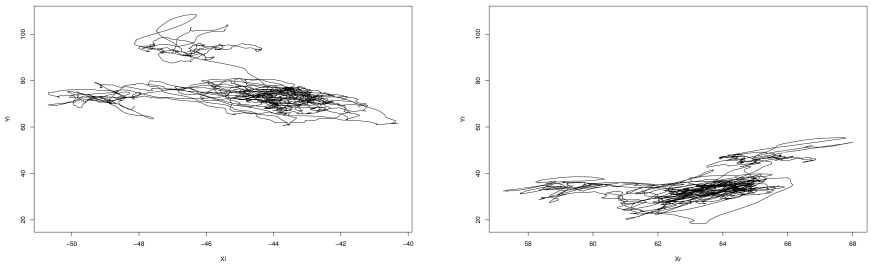


Figure 1: Représentations des suites $t \mapsto L_t$ (à gauche) et $t \mapsto R_t$ (à droite) des positions du centre de pression sur la période T des deux pieds sur la plateforme de force, issues du protocole 3 (voir Table 1) associée à un sujet sain.

résumé des données brutes en ne considérant que la suite $(C_t)_{t\in T}$ des distances séparant la trajectoire de la projection du centre de gravité du patient sur la plateforme de force à un point de référence b. Pour cela nous introduisons la suite composée du milieux des segments: $(B_t)_{t\in T} = (\frac{1}{2}(L_t + R_t))_{t\in T}$. Nous définissons b comme la valeur médiane de $(B_t)_{t\in T}$ durant la première phase du protocole. Ainsi, $(C_t)_{t\in T}$ est défini comme suit: $C_t = \|B_t - b\|_2$ pour tout $t \in T$. En Figure 2 nous donnons une représentation visuelle de $(C_t)_{t\in T}$ issue de deux protocoles associés à un sujet hémiplégique.

Puisque sans doute, c'est au voisinage de la phase de perturbation d'un protocole que se situe une grande partie des caractéristiques d'une trajectoire (voire Figure 2), nous proposons de ne considérer comme mesures résumées de $(X_t)_{t\in T}$ qu'un vecteur Y de dimension finie défini comme suit:

$$Y = (\bar{C}_1^+ - \bar{C}_1^-, \bar{C}_2^- - \bar{C}_1^+, \bar{C}_2^+ - \bar{C}_2^-) \tag{1}$$

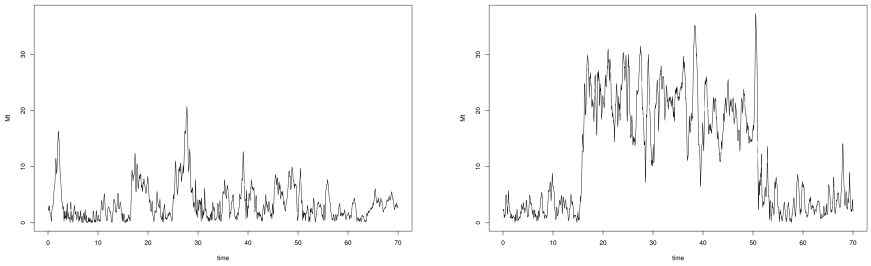


Figure 2: Représentation d'une trajectoire $t \mapsto C_t$ qui correspond à deux protocoles associés à un patient hémiplégique (premier protocole à gauche, troisième protocole à droite).

où

$$\bar{C}_1^- = \frac{\delta}{5} \sum_{t \in T \cap [10,15[} C_t, \qquad \bar{C}_1^+ = \frac{\delta}{5} \sum_{t \in T \cap]15,20]} C_t,$$

$$\bar{C}_2^- = \frac{\delta}{5} \sum_{t \in T \cap [45,50[} C_t, \qquad \bar{C}_2^+ = \frac{\delta}{5} \sum_{t \in T \cap]50,55]} C_t$$

sont les moyennes de C_t calculées sur les intervalles $[10, 15[$, $]15, 20]$, $[45, 50[$ et $]50, 55]$ (ce qui représentent les dernières/premières 5 secondes avant/après le début/fin de la seconde phase du protocole d'intérêt).

3 Cadre et objectif statistique

La structure d'une donnée observée O s'écrit comme: $O = (W, A, Y^1, Y^2, Y^3, Y^4)$, où

- $W \in \mathbb{R} \times \{0, 1\}^2 \times \mathbb{R}^2$ est le vecteur des covariables (voir Section 2);
- $A \in \{0, 1\}$ indique l'état du patient (avec la convention $A = 1$ pour un patient hémiplégique et $A = 0$ pour un patient sain);
- pour tout $j \in \{1, 2, 3, 4\}$, $Y^j \in \mathbb{R}^3$ est la mesure résumée (comme défini par (1)) associée au j-ième protocole.

On note P_0 la loi de O où $P_0 \in \mathcal{M}$ avec $\mathcal{M}$ l'ensemble de toutes les lois possibles pour O. Notre objectif principal est ici de construire un classifieur ϕ basé sur W et (Y^1, Y^2, Y^3, Y^4) pour déterminer quand $A = 0$ ou $A = 1$. Afin d'étudier quels sont les protocoles les plus adaptés au trouble de la posture relatif à l'hémiplégie, nous proposons le classement des protocoles (du plus informatif au moins informatif) puis nous construisons

quatre classifieurs $\phi^1, \phi^2, \phi^3, \phi^4$ qui sont basés respectivement sur le meilleur protocole, les deux meilleurs, les trois meilleurs et les quatre protocoles. Pour le classement des protocoles nous proposons de nous appuyer sur le critère Ψ suivant: $\Psi : \mathcal{M} \to \mathbb{R}^{12}$ tel que, pour tout $P \in \mathcal{M}$, $\Psi(P) = (\Psi^j(P))_{j\leq 4}$ où

$$\Psi^j(P) = \left(E_P\left\{E_P[Y_i^j|A=1,W] - E_P[Y_i^j|A=0,W]\right\}\right)_{1\leq i\leq 3}.$$

La composante $\Psi_i^j(P)$ est connue dans la littérature comme la mesure de l'importance de la variable A sur la mesure résumée Y_i^j conditionnellement à W (voir Van der Laan et Robins (2003)).

4 Méthode mise en place

Le classement des protocoles que nous proposons est basé sur le test d'hypothèse nulle suivante:

$$\text{“}\Psi_i^j(P_0) = 0\text{”}, \quad (i,j) \in \{1,2,3\} \times \{1,2,3,4\}.$$

Heuristiquement, rejeter “$\Psi_i^j(P_0) = 0$” signifie que la valeur de la i-ième coordonnée de Y^j fournit beaucoup d'informations pour déterminer si $A = 0$ ou $A = 1$. Nous nous appuyons sur une méthode robuste, la méthodologie du maximum de vraisemblance ciblé (voir Van der Laan et Rubin (2006)). Soient $O_{(1)}, \ldots, O_{(n)}$ n copies indépendantes de O. Pour tout $(i,j) \in \{1,2,3\} \times \{1,2,3,4\}$, on calcule l'estimateur du maximum de vraisemblance ciblé (TMLE) $\Psi_{i,n}^j$ de Ψ_i^j basé sur $O_{(1)}, \ldots, O_{(n)}$ et un estimateur $\sigma_{i,n}^j$ de sa variance asymptotique. Ceci conduit alors à considérer la t-statistique $T_{i,n}^j = \frac{\sqrt{n}\Psi_{i,n}^j}{\sigma_{i,n}^j}$ pour tout $(i,j) \in \{1,2,3\} \times \{1,2,3,4\}$. Notre classement des protocoles s'appuie donc sur $(T_{1,n}^j, T_{2,n}^j, T_{3,n}^j)$. Nous décidons que le protocole j est plus informatif que le protocole j' si

$$\sum_{i=1}^{3}(T_{i,n}^{j'})^2 < \sum_{i=1}^{3}(T_{i,n}^{j})^2.$$

La construction des classifieurs ϕ^j est basée sur le super learning (voir Van der Laan, Polley et Hubbard (2007)), qui est une méthode d'agrégation de classifieurs fondée sur la validation croisée. Étant données une librairie de prédicteurs et une fonction de perte (L), le super learning consiste à sélectionner par validation croisée, la combinaison linéaire convexe de prédicteurs de la librairie qui minimise le risque induit par L.

5 Résultats

Nous évaluons les performances des classifieurs ϕ^j sur les données réelles en utilisant la règle du leave-one-out. La performance du classifieur ϕ^j est notée Perf^j. Les résultats sont

données en Table 2. Notre librairie de prédicteurs s'appuie notamment sur des méthodes de classification telles que les forêts aléatoires (voir Breiman (2001)) ou encore le top scoring pairs classifier (voir Geman, d'Avignon, Naiman, Winslow et Zeboulon (2004)).

	$j=1$	$j=2$	$j=3$	$j=4$
Perf^j	0.74	0.83	0.82	0.87

Table 2: Performance des classifieurs ϕ^j, $j \in \{1,2,3,4\}$, Perf^j représente le taux de patients bien classés.

La Table 2 montre que l'usage des deux protocoles les plus informatifs donne des résultats de classification satisfaisants (83%). La description des protocoles en Table 1 nous indique que les patients hémiplégiques (qui souffrent de troubles proprioceptifs) semblent privilégier le système visuel. Ceci est en accord avec la littérature médicale (voir Bonan, Yelnik, Laffont, Vitte et Freyss (1996)). Néanmoins la meilleure performance de classification est obtenue avec les quatre protocoles, le taux de bien classés est alors de 87%. Ce taux est déjà très satisfaisant étant donnée la complexité du problème. On peut espérer l'améliorer en considérant une autre mesure résumée Y. Par exemple, on peut penser à des caractéristiques d'un modèle de diffusion pour $(C_t)_{t\in T}$.

Bibliographie

[1] Bonan, I. Yelnik, A. Laffont, I. Vitte, E. et Freyss, G. (1996) *Selection of sensory information in postural control of hemiplegics after unique stroke*, Annales de Réadaption et de Médecine Physique, 39, 157-163.

[2] Breiman, L. (2001) *Random Forests*, Machine Learning, 45, 5-32.

[3] Chambaz, A. Bonan, I. et Vidal, P-P. (2009) *Deux modèles de Markov cachés pour processus multiples et leur contribution à l'élaboration d'une notion de style posturale*, Journal de la Société Française de Statistique, 150(1), 73-100.

[4] Geman, D. d'Avignon, C. Naiman, D. Winslow, R. et Zeboulon, A. (2004) *Classifying gene expression profiles from pairwise mRNA comparisons*, Statistical Applications in Genetics and Molecular Biology, 3, Art.19.

[5] Van der Laan, M. J. Polley, E. C. et Hubbard, A. E. (2007) *Super learner*, Statistical Applications in Genetics and Molecular Biology, 6, Art.25.

[6] Van der Laan, M. et Robins, J. M. (2003) *Unified methods for censored longitudinal data and causality*, Springer.

[7] Van der Laan, M. et Rubin, D. (2006) *Targeted maximum likelihood learning*, The International Journal of Biostatistics, 2, Art.11.

Une 2D-réduction de dimension par un estimateur de la distance en probabilité de Patrick-Fisher

WISSAL DRIRA & FAOUZI GHORBEL

Laboratoire CRISTAL, pôle GRIFT, Ecole National des Sciences de l'Informatique
Campus Universitaire de la Manouba, 2010 Manouba, Tunisie
Wissal.drira@gmail.com, Faouzi.ghorbel@ensi.rnu.tn

Abstract

Here, we intend to evaluate the performance of the 2D-dimension reduction algorithm obtained by a novel estimator of Patrick-Fisher distance based upon the orthogonal series. Such algorithm is compared to Fisher discriminate analysis in the mean of the probability of error which we propose to estimate with the 2D-Kernel probability density function estimate.

Résumé

Dans cet article, nous proposons d'évaluer les performances de la 2D-réduction de dimension par un nouvel estimateur de la distance de Patrick-Fisher à l'aide des fonctions orthogonales et cela en estimant les probabilités d'erreur. La réduction de dimension proposée est comparée avec l'analyse discriminante de Fisher au sens de la probabilité d'erreur laquelle est estimée à l'aide de l'estimateur à noyaux dans l'espace réduit.

Mots clés: Réduction de dimension, distance de Patrick-Fisher, analyse discriminante de Fisher, méthode du noyau, probabilité d'erreur de classification

1. Introduction

Un des problèmes les plus étudiés et essentiels dans le domaine de la reconnaissance de formes statistique est la sélection de primitives discriminantes. L'ensemble de ces primitives est souvent représenté par des vecteurs aléatoires de grande dimension D. Ces vecteurs aléatoires sont souvent supposés absolument continus relativement à la mesure de Lebesgue de R^D. Les densités de probabilité conditionnelles aux classes doivent être estimées afin d'appliquer la règle de classification de Bayes. L'estimation de ces densités de probabilité en grande dimension nécessite des tailles d'échantillon très grandes. En effet, pour une précision donnée, les théorèmes de convergence des estimateurs des densités de probabilité requièrent des tailles d'échantillons qui augmentent exponentiellement avec la dimension D des vecteurs aléatoires. La d-réduction de dimension permet de ramener l'estimation dans l'espace réduit de dimension d<<D.

Il est bien connu que la réduction de dimension par la méthode des matrices de dispersion donne des résultats satisfaisants au sens de la complexité, de la simplicité et de la précision. Toutefois, lorsque par exemple une des densités de probabilité conditionnelle est multimodale, le critère de Fisher n'aboutit pas à une réduction satisfaisante puisque le critère de séparabilité associé ne fait intervenir que les moments statistiques d'ordre inférieur ou égal à 2.

D'autres méthodes optimisant les quantités comme les distances en probabilité ou l'information mutuelle [2] ont été introduites pour palier à cet inconvénient. La mise en œuvre de la distance de Patrick-Fisher n'aboutit qu'à un extracteur scalaire. Au sens de la probabilité d'erreur, A. Hillion et al ont montré la meilleure tenue de l'analyse discriminante obtenue par la distance de Patrick-Fisher vis-à-vis de celle se basant sur les matrices de dispersion [3]. Toutefois, son extension à la d-réduction (d>1) ne semble pas être aisée pour l'estimateur introduit initialement par Patrick et Fisher. Des méthodes récurrentes ont été proposées dans la littérature pour donner une première réponse à ce problème [5].

Dans un précédant article, nous avons proposé d'introduire un nouvel estimateur pour la distance de Patrick-Fisher basé sur les fonctions orthogonales déjà utilisées pour l'estimation des densités de probabilité. Une telle méthode permet la réduction de dimension vers un espace réduit de dimension supérieur à 2. Dans ce sens, nous proposons d'étendre la formulation de l'extracteur scalaire introduit dans [6] au cas de la réduction de dimension vers un plan discriminant. L'évaluation au sens de la probabilité d'erreur est réalisée par une estimation basée sur la méthode à noyau. La 2D-réduction proposée est comparée à celle de l'analyse discriminante de Fisher au sens du taux de mauvaise classification.

2. Analyse discriminante de Fisher

La méthode de réduction de dimension connue sous le nom d'analyse discriminante de Fisher a été introduite pour le cas de deux classes puis généralisée par Rao au cas de plusieurs classes. Une matrice rectangulaire W représentant l'extraction de primitives de l'espace des primitives de dimension D vers un espace réduit de dimension d($d \ll D$). Elle est obtenue par la maximisation du critère de Fisher qui est défini comme le rapport des traces des estimateurs des matrices de dispersion exprimées dans l'espace réduit :

$$J(w) = \frac{trace(w\hat{S}_b w^T)}{trace(w\hat{S}_w w^T)}$$

Où $\hat{S}_b \; et \; \hat{S}_w$ sont respectivement les estimateurs des matrices inter classe S_b et intra classe S_w. Nous rappelons ici leurs expressions respectives:

$$\hat{S}_b = \sum_{k=1}^{K} \pi_k (\mu_k - \mu)(\mu_k - \mu)^t \quad \text{et} \quad \hat{S}_w = \sum_{k=1}^{K} \pi_k \, \frac{1}{N_k} \sum_{i=1}^{N_k} (X_i^k - \mu_k)(X_i^k - \mu_k)^t$$

$$\text{avec } \mu = \sum_{k=1}^{K} 1/_{N_k} \sum_{i=1}^{N_k} X_i^k \quad \text{et } \mu_k = 1/_{N_k} \sum_{i=1}^{N} X_i^k$$

Où $\{X_i^k, i = 1, .., N\}$, représente un échantillon d'apprentissage supervisé de taille N. k désigne la classe du vecteur X_i^k de dimension D.

La matrice rectangulaire correspondante à W est formée par les d vecteurs propres associées aux d plus grandes valeurs propres de la matrice $(\hat{S}_b)(\hat{S}_w)^{-1}$. Comme ces matrices de dispersion sont définies seulement à partir des moments statiques d'ordre inférieur ou égal à 2. Par conséquent, dans des situations un peu complexes telles que la multimodalité de la distribution de l'observation relative à une même classe, les moments d'ordre faibles ne peuvent pas décrire complètement sa dispersion statistique. Dans ce contexte l'analyse discriminante basée sur le critère de Fisher n'est plus capable de donner satisfaction. Dans l'objectif de s'affranchir de cette limitation, des distances entre les densités de probabilité conditionnelles pondérées par les probabilités a priori ont été suggérées dans la littérature. Dans ce sens, nous nous intéressons à la distance de Patrick-Fisher qui s'écrit comme suit :

$$\mathrm{d}_2(\pi_1 \mathrm{f}_1, \pi_2 \mathrm{f}_2) = \left(\int_{R^D} |\pi_1 \mathrm{f}_1 - \pi_2 \mathrm{f}_2|^2 \mathrm{dx} \right)^{1/2}$$

Où f_k représente la densité de probabilité conditionnelle à la classe k et π_k la probabilité a priori associée. Dans le cadre de la classification d'images de texture, A. Hillion et al ont proposé d'étudier les performances de la réduction de dimension obtenue par une telle distance en évaluant les probabilités d'erreur [3]. L'estimateur de cette distance introduite par Patrick et Fisher, est obtenu en remplaçant dans l'expression de la distance les densités de probabilité conditionnelles par leur estimateur à noyau. En considérant le noyau gaussien, l'expression de la distance devient une somme finie de fonctions de Gauss.

3. Un 2D-extracteur linéaire basé sur un nouvel estimateur de la distance de Patrick-Fisher

Dans le présent paragraphe, nous décrivons l'extension de l'extracteur scalaire basé sur des fonctions orthogonales que nous avons introduit dans [6] au cas de la sélection de primitives bidimensionnel (i.e. d=2). Cet estimateur suppose que les densités de probabilité conditionnelles appartiennent à l'espace $L^2(\mathbf{U})$ avec U un rectangle du plan réel R^2. Cet espace de Hilbert peut être muni d'une base de fonctions orthogonales que nous notons par:

$$e_{l,k}(x,y)$$

Où k et l sont des entiers. Souvent, il est possible d'obtenir cette base par le produit des fonctions de base unidimensionnelles. Ainsi l'estimateur d'une densité de probabilité d'un couple aléatoire (X,Y), s'écrit :

$$\hat{f}_{(X,Y)}(x,y) = \sum_{l=1}^{m_N}\sum_{k=1}^{m_N} \hat{a}_{l,k}\, e_{l,k}(x,y)$$

Où m_N est appelé paramètre de troncature. La densité conjointe du couple peut s'écrire aussi :

$$\hat{f}_{(X,Y)}(x,y) = \sum_{l=1}^{m_N}\sum_{k=1}^{m_N} \hat{a}_{l,k}\, e_l{}^*(x)\, e_k(y)$$

Où $\hat{a}_{l,k}$ est l'estimateur du coefficient de Fourier $a_{l,k}$ qui s'exprime de la manière suivante:

$$\hat{a}_{l,k} = \frac{1}{N}\sum_{i=1}^{m_N} e_l{}^*(X_i)\, e_k(Y_i)$$

m_N joue ici le rôle d'un facteur de lissage. La convergence en moyenne quadratique intégrée de cet estimateur est obtenue quand la suite m_N est équivalente $N^{-1/s}$ avec $s > 2$.

L'estimateur des fonctions orthogonales peut s'écrire aussi comme une sommation finie de noyau généralisant ainsi de la méthode du noyau. :

$$\hat{f}_{(X,Y)}(x,y) = \frac{1}{N}\sum_{i=1}^{N} K'_{m_N}(x,y,X_i,Y_i)$$

Avec :

$$K'_{m_N}(x,y,X_i,Y_i) = K_{m_N}(x,X_i).K_{m_N}(y,Y_i)$$

Où $K_{m_N}(x,X_j)$ est le noyau généralisé pour une base unidimensionnelle de fonctions orthogonales. Sachant que K'_{m_N} vérifie les même propriétés que K_{m_N} détaillées dans [6]. En remplaçant dans l'expression de la distance de Patrick-Fisher bidimensionnel toutes les quantités par les estimateurs correspondant et en utilisant l'orthogonalité des fonctions de la base, nous aboutissons à un estimateur dont l'expression élevée au carrée est donnée par :

$$\hat{d}_2{}^2(f_1,f_2) = \frac{1}{N^2}\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\left[K'_{m_1}(X_i,Y_i,X_j,Y_j)\right] + \sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\left[K'_{m_2}(X_i,Y_i,X_j,Y_j)\right] - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left[K'_{\min(m_1,m_2)}(X_i,Y_i,X_j,Y_j)\right]\right)$$

Ainsi $\hat{d}_2$ représente une extension de l'estimateur unidimensionnel introduit [6].

4. Evaluation des performances

Il est bien connu que le meilleur critère d'évaluation de performance d'une méthode de réduction de dimension est la probabilité d'erreur. Dans le cas unidimensionnel, l'estimation de la probabilité d'erreur de Bayes revient souvent à l'approximation des aires de recouvrement entre les densités de probabilité conditionnelles pondérées par les probabilités a priori. Dans le cas bivarié, l'extension de cette dernière méthode d'estimation requière l'approximation des volumes de recouvrement entre les surfaces des densités de probabilité conditionnelles. Cette tâche n'est pas aisée numériquement et est à l'origine d'erreurs pouvant dépasser dans la plupart des cas la valeur de la probabilité d'erreur recherchée. Pour cela nous proposons de procéder par la série d'étapes suivantes:

Etape 1 : La génération d'un échantillon d'apprentissage servant à trouver la transformation optimale W qui représente la 2D réduction de dimension.

Etape 2 : Le calcul de l'extracteur vectoriel W par l'optimisation de l'estimateur de la distance de Patrick-Fisher proposée entre les classes.

Etape 3 : La mise en œuvre de la projection de cet échantillon sur le plan engendré par la transformation W .

Etape 4 : La génération d'un échantillon test ayant pour objectif l'estimation de la probabilité d'erreur de Bayes.

Etape 5 : Le taux de mauvaise classification est ainsi calculé en chaque valeur d'observation de l'échantillon test projeté sur W et cela grâce à la comparaison des valeurs des estimations des densités de probabilité conditionnelles pondérées par les proportions de chaque classe en ces observations tests. Ces densités de probabilité sont estimées à l'aide de la méthode des noyaux en se basant sur l'échantillon d'apprentissage.

Dans le tableau 1, nous donnons le taux de mauvaise classification pour les réductions obtenues respectivement par la méthode de Fisher basée sur les matrices de dispersion et par l'optimisation de l'estimateur de la distance de Patrick-Fisher.

Tab. 1. Taux de mauvaise classification

D dimension de l'espace de départ	Taille de l'échantillon d'apprentissage $N_1 = N_2$	Taille de l'échantillon test N_1+N_2	Méthode de Fisher	Méthode de Patrick-Fisher
5	1000	500 + 500	0.0020	0.0020
		1000 + 1000	0.0020	0.0070
	2000	2000 + 1800	0.0026	0.0076
3	1000	100 + 400	0.0130	0.0125
		200 + 900	0.0687	0.0215
		850 + 650	0.0152	0.0157
		950 + 900	0.0257	0.0122
	2000	1500 + 1500	0.5630 (*)	0.0052
		2000 + 1800	0.0263	0.0153

(*) Ce résultat est expliqué par le fait que l'inversion numérique de $\hat{S}_w$ n'a pas été effectuée correctement. Dans ce cas nous avons pu vérifier que la matrice à inverser était mal conditionnée.

La meilleure performance de la distance de Patrick-Fisher en analyse discriminante relativement à celle obtenue par les matrices de dispersion est observée par un taux de mauvaise classification inférieur ou égale (tab. 1). Nous remarquons que les performances sont les mêmes lorsque les lois conditionnelles sont gaussiennes et de vecteurs moyens relativement distants par rapport aux valeurs des traces des matrices de covariance de chaque classe. Dans les cas où une classe est entourée ou enveloppée par la seconde, le plan discriminant obtenu par la distance de Patrick-Fisher est plus séparateur au sens du taux de mauvaise classification.

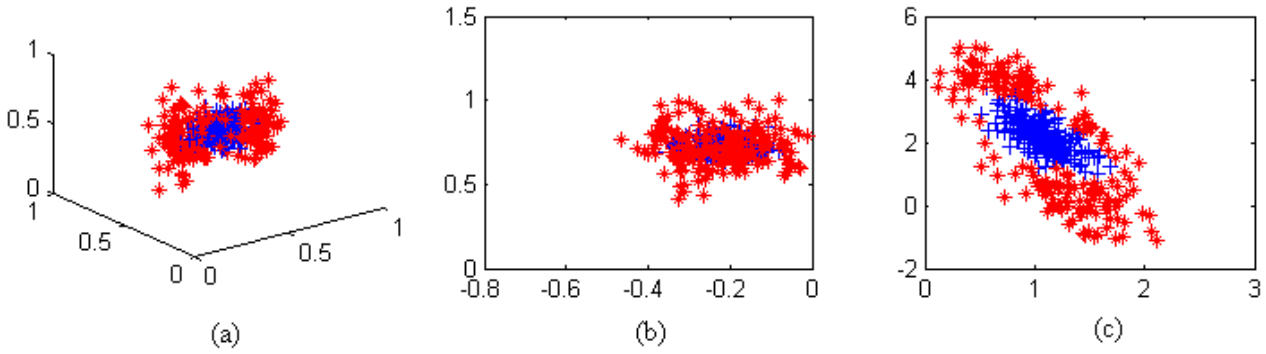


Figure 1. (a) Le nuage d'un mélange gaussien- non gaussien dans l'espace d'origine. (b) Le résultat obtenu par la méthode de Fisher. (c) Le résultat de réduction de dimension obtenu par l'estimateur proposé.

La figure 1(a) représente les deux nuages des observations de chaque classe dans l'espace initial $(\mathrm{D} = 3, \mathrm{X} \in \mathbb{R}^3)$ générés selon un mélange de deux classes équiprobable avec un vecteur gaussien pour la première classe, et un vecteur gaussien tronqué suivant une boule centrée à l'origine pour la seconde classe. La figure 1(b) représente le nuage projeté sur le plan discriminant obtenu par la méthode de Fisher. La figure 1(c) représente le nuage projeté sur le plan discriminant obtenu par la méthode proposée. Visuellement, nous observons une meilleure tenue de la méthode proposée.

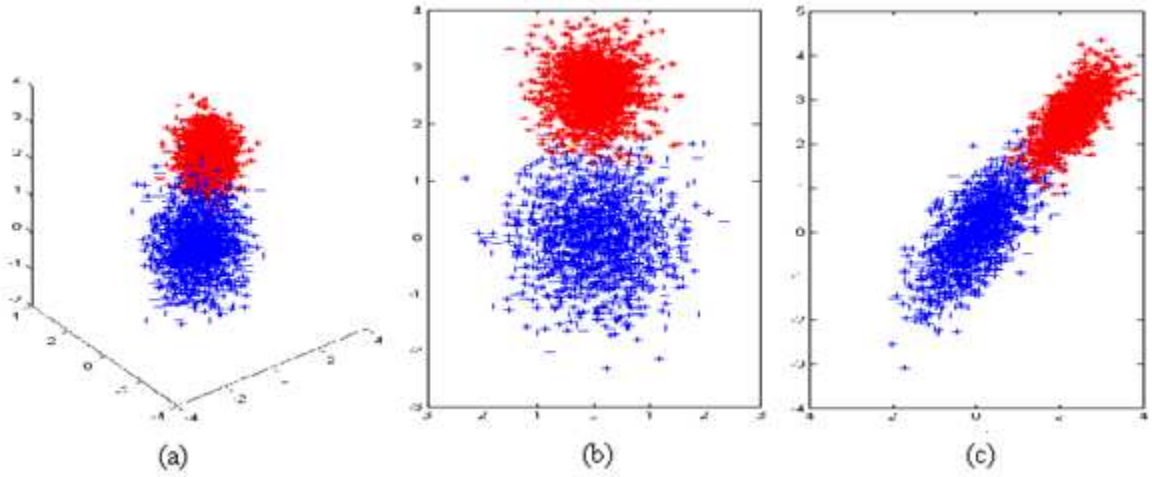


Figure 2. (a) Le nuage gaussien dans l'espace d'origine avant réduction de dimension. (b) Le résultat obtenu par la méthode de Fisher. (c) Le résultat de réduction de dimension obtenu par l'estimateur proposé.

La figure 2(a) représente le nuage d'observations d'un cas Gaussien-hétéroscedastic dans l'espace initial $(\mathrm{D} = 3, \mathrm{X} \in \mathbb{R}^3)$. Le nuage projeté sur le plan discriminant obtenu par la méthode de Fisher est représenté dans la figure 2(b). La figure 2(c) visualise le même nuage projeté sur le plan discriminant obtenu par l'estimateur proposé de la distance de Patrick-Fisher.

Il ressort des illustrations des figures 2(a), (b) et (c) que la distance de Patrick-Fisher fournit un plan discriminant comparable à celui de Fisher. Le tableau tend cependant à montrer qu'on observe une meilleure stabilité et une relative meilleure performance au sens du taux de mauvaise classification.

5. Conclusion

Dans cet article, nous avons introduit un estimateur de la distance de Patrick-Fisher par les fonctions orthogonales qui permet de réaliser une 2D-réduction de dimension. Cela est devenu possible grâce à la définition de noyaux produits. Ainsi, cela nous a permis de tester les performances de l'analyse discriminante basée sur la distance Patrick-Fisher dans le cas de la réduction de dimension vers un espace bidimensionnel. L'évaluation des performances précise au sens de la probabilité d'erreur est devenue possible grâce à l'estimateur de la probabilité d'erreur par la méthode du noyau.

En perspective de ce travail une étude sur l'analyse discriminante tenant compte d'information sur le support des données à classifier par ces deux estimateurs sera menée. En effet, pour chaque type d'espace de Hilbert, nous disposons de bases de fonctions adaptées comme ceux de Legendre, Laguerre, Hermite ...

L'extension au cas du multi classe fera l'objet de nos travaux futurs. Son application aux données réelles représentera un enjeu important.

Bibliographie

[1] Z. Nenadic, "Information Discriminant Analysis: Feature Extraction with an Information-Theoric Objective", IEEE transactions on PAMI, Vol 29 N° 8 August 2007

[2] E.A. Patrick and F.P. Fisher, "Non parametric feature selection", IEEE Trans. On Inf. Theory, vol. IT-15, pp.577-584, 1969.

[3] A. Hillion and al. "A non parametric approach to linear feature extraction; Application to classification of binary synthetic textures", 9th International Conference on Pattern Recognition, ICPR pp.1036-39, 1988.

[4] B. W Silverman, "Density Estimation for Statistics and Data Analysis", London, Charman and Hall, 1986.

[5] M.E. Aladjem, "PNM: A program for parametric and nonparametric mapping of multidimensional data", Computers in Biology and Medicine, vol. 21,pp. 321-343, 1991.

[6] F. Ghorbel et W. Drira, "Réduction de dimension par un nouvel estimateur de la distance de Patrick Fisher à l'aide des fonctions orthogonales", 42èmes Journées de Statistique, Marseille 2010.

[7] M.E. Aladjem, "Two class pattern discrimination via recursive optimization of Patrick-Fisher distance", Proc. of the 13th International Conference on Pattern Recognition, vol. 2, pp.60-64, 1996.

[8] M.E. Aladjem, "Nonparametric discriminant analysis applied to medical diagnosis", Proc. of the 19th Convention of Electrical and Electronics Engineers in Israel, 1996 (in press).

[9] M.E. Aladjem, "Linear discriminant analysis for two-classes via removal of classification structure", IEEE Trans. Pattern Anal. Mach. Intell., 1997.

[10] P. Diaconis and D.Freedman, "Asymptotics of graphical projection pursuit", The Annals of Statistics, vol.12, pp.793-815, 1984.

[11] Marco Loog and al, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria", IEEE transaction on PAMI Vol 23 N° 7 Juilly 2001

Robust Adaboost for data fusion problems

Afef Ben brahim & Mohamed Limam

LARODEC, ISG, University of Tunis
41 Avenue de la liberté, cité Bouchoucha, 2000 Le Bardo, Tunisia

afef.benbrahim@yahoo.fr
mohamed.limam@isg.rnu.tn

Résumé

Dans plusieurs problèmes de classification, les données sont générées à partir de différentes sources. L'exploitation de toutes les données disponibles est importante pour une prise de décision efficace. La fusion de sources de données hétérogènes pour un même problème est naturellement adaptée aux méthodes d'ensemble puisque différents classifieurs peuvent être construits en utilisant des données obtenues à partir de multiples sources, puis combinés pour avoir un résultat final. Nous proposons des méthodes robustes pour la combinaison de classifieurs, ayant pour objectif de réduire l'effet des classifieurs non fiables dans l'ensemble. Les méthodes proposées montrent une amélioration des résultats de classification.

Mots clés: Classification; fusion de données; combinaison de classifieurs; résolution de conflit.

Abstract

In many classification problems, data are generated from different sources and views. Taking advantage of all the data available is important for intelligent decision making. Fusion of heterogeneous data sources underlying the same problem presents a natural fit for ensemble systems since different classifiers could be generated using data obtained from different sources, and then combined to achieve the desired data fusion. Robust methods are proposed for combining classifiers, aimed at reducing the effect of outlier classifiers in the ensemble. The proposed methods are shown to have better performance leading to significantly better classification results than the previously employed techniques.

Keywords: Classification; data fusion; classifiers combination; conflict resolution.

1 Introduction

The idea of Ensemble methods is to construct a set of classifiers, such as decision trees or neural networks, for the same original problem. To classify a new instance, decisions of single classifiers are combined by voting or averaging leading to a more accurate classification decision.

Diversity is a key element for the effectiveness of ensemble methods. There are different strategies to achieve diversity based mostly on the manipulation of the learning data by techniques such as sampling, partitioning.

Fusion of heterogeneous data sources presents a natural fit for ensemble systems since different classifiers could be generated using data obtained from different sources. This paper addresses the issue of how to effectively use ensemble methods to optimally combine multiple sources of information in order to make a decision. An ensemble method, Adaboost.M1 proposed by Freund and Schapire (1997), is evaluated for data fusion problems. Moreover, the use of robust combination rules is investigated. Our proposed methods aim at reducing the effect of outlier classifiers in the ensemble. The algorithm and the variations obtained by the use of different combination methods are evaluated experimentally on five benchmark databases.

2 Ensemble methods for data fusion

Many sources of data can present different related views of the same phenomenon. This results in challenging machine learning problems where data sources are combined in order to benefit of the complementary information brought by each source. This process is known by data fusion.

Data fusion processes are categorized into three main levels, low, intermediate and high, depending on the stage at which fusion is performed. Several methods of decision fusion exist, such as statistical methods, voting methods and ensemble methods where diversity is an important requirement for constructing good ensemble of classifiers. This diversity is precisely achieved by using different sources of training data. Several ensemble approaches have been proposed for data fusion problem. Briem et al (2002) experiment Bagging, Boosting, and consensus-theoretic classifiers for the classification of multisource remote sensing and geographic data.

Polikar et al (2008) developed Learn++ which is another ensemble approach for data fusion applications. This algorithm was originally developed for incremental learning of novel information from new data and adopted for data fusion.

Adaboost.M1 (Freund and Schapire (1997)) is one of the most popular and effective ensemble approaches. Thus, we decided to test its performance with our proposed combination methods.

Adaboost.M1 generates an ensemble of hypotheses by training a weak classifier, and combines their outputs using weighted majority voting (WMV). Iteratively, the individual classifiers are trained with instances drawn after updating the distribution of the learning data. This distribution update is based on the performance of the previous classifier on the training data forcing the algorithm to focus on instances that have been misclassified. Adaboost.M1 requires a base learning algorithm which have an error less than 0.5.

The idea of WMV is to assign a weight to each classifier proportionally to its estimated performance and the final decision is influenced by classifiers having a high estimated perfor-

mance. So, the quality of the classifier is important in this group decision making process, and it depends only on the model's own estimated performance. However, the classifier with the largest weight could be unreliable.
There is a need to search for more robust aggregation rules allowing to achieve the desired data fusion and obtain better classification performance.

We propose to use a robust combination rule for aggregating classifiers, aimed at reducing the effect of outlier classifiers in the ensemble. The proposed solution takes into account the conflict level of a classifier with the other classifiers in the group.

3 Data fusion by robust classifiers combination

This technique aims at reducing the influence of conflicting classifiers in the ensemble (Garcia and Puig (2004)). The process involves two steps. The first one is the training stage where an ensemble of classifiers is trained on the learning dataset, the second one is the conflict resolution and decision making.

3.1 Training stage

Let $O_j = \{o_{tj}, t = 1 \dots T\}$ be the opinions of an ensemble of T classifiers, regarding a set of C classes, $\Omega = \{\omega_j, j = 1 \dots C\}$, corresponding to a classification problem. A confidence level w_{tj} is assigned to each classifier about each opinion o_{tj} it expresses. The confidence is a weight based on the Kullback J-divergence (KJ) (kittler (1986)) measuring the separability between two classes ω_a and ω_b as follows

$$KJ_t(\omega_a, \omega_b) = \int_0^1 (A - B) \log(\frac{A}{B}) du, \tag{1}$$

where A and B are obtained from classifier's opinions regarding respectively the two classes ω_a and ω_b. This method measures the classifier's confidence. A classifier with low KJ measure will have a low confidence, as it separates slightly the different classes. A classifier with high KJ differentiates properly among the different classes. Such classifier will have a high confidence. These confidences are computed as the normalized average of the KJ between ω_j and the other classes. This is a possible way to define expert's confidence.
Another confidence formulation is to use weights given by

$$w_{tj} = \log(\frac{1}{\beta_t}), \tag{2}$$

where β_t is the normalized error of the t^{th} classifier in the ensemble for class ω_j.

3.2 Conflict resolution and decision making

Given $O_j(x)$ the opinions of T classifiers about the belonging of an instance x to the class ω_j, and given $W_j = \{w_{tj}, t = 1 \dots T\}$, the confidences associated with those opinions, the conflict of each classifier is formulated by first measuring the similarity between its opinions and those of the other classifiers in the ensemble as follows

$$Sim_t(O_j(x)) = 1 - \frac{1}{(T-1)} \sum_{k=1,k\neq t}^{T} \mid o_{tj}(x) - o_{kj}(x) \mid . \tag{3}$$

Then, expert's confidence similarity with the rest of confidences, $Sim_t(W_j)$, is calculated the same way as in Eq. (5). Based on these calculations, the conflict raised by a classifier is defined as

$$Conflict_{tj}(x) = Sim_t(W_j)[1 - Sim_t(O_j(x))]. \tag{4}$$

Conflicting classifiers are those with similar confidences to the agreeing classifiers but completely different opinions from theirs. The conflict measure will affect classifier's reliability which is calculated as follows

$$r_{tj}(x) = w_{tj}(1 - Conflict_{tj}(x)). \tag{5}$$

Finally, the original opinions of the experts are adjusted by multiplying them by the associated reliability factors after being normalized. The selected class is the one having the maximum adjusted opinion.

Given that for each classifier two confidence formulations are possible, two versions of the robust combination method, namely robust combination based on KJ divergence criterion denoted by RKJ and robust combination based on $\log(\frac{1}{\beta})$ criterion denoted by RL, could be used.

4 Experimental setup and results

4.1 Proposed algorithms

The robust aggregation rule is evaluated on Adaboost.M1. It is used in the data fusion process to integrate the classifier ensembles of all feature sets. Two variations of Adaboost.M1, namely Adaboost.RKJ and Adaboost.RL, which use respectively RKJ and RL as combination techniques, are obtained. Ensembles of 10 classifiers, using Multilayer perceptrons (MLP) and decision trees (DT) as base learners, are evaluated for each feature set, repeating this process 10 times in order to get an average estimate of the performance.

4.2 Results on Multiple features database

This database is obtained from the UCI Machine Learning Repository and it is represented by 6 feature sets. Individual and fusion performances of Adaboost.M1 and its two variations Adaboost.RKJ and Adaboost.RL are shown in Table (1).

Table 1: Fusion performances obtained by Adaboost algorithm with MLP and DT on multiple features dataset.

Feature set / Classifier	MLP	DT
FS1	**90.53**	80.35
FS2	71.47	70.18
FS3	86.71	73.29
FS4	85.00	81.62
FS5	74.06	71.15
FS6	73.94	69.89
Adaboost.M1	96.83	96.09
Adaboost.RKJ	96.95	**97.14**
Adaboost.RL	97.02	91.14

MLP always outperforms DT. The best classification accuracy is obtained by FS1 with an ensemble of 10 MLP classifiers. As expected, fusion results exceed individual ones with best performance obtained by Adaboost.RKJ with an ensemble of 10 DT classifiers, leading to an improvement of about 7% compared to FS1 result.

4.3 Other databases

This section summarizes evaluation results of Adaboost.M1 and its proposed variations for fusing multiple feature sets of four other benchmark databases. These databases are randomly partitioned into subsets, where each partition uses only a portion of the features to simulate a data fusion setting. Results provided in Table (2) are the best individual and fusion results obtained by each algorithm for each database when comparing MLP and DT results. Empty boxes in Table (2) indicate that fusion results are lower than best individual feature set result.

Fusion classification results are often better than individual results. Generally, Adaboost.RL gives the best results and can be considered the most robust as it outperforms the best individual classification accuracy for all databases. Adaboost.RKJ, is also effective and gives the best fusion results for two databases, Multiple feature and Wine. However, empty box shows that it gives lower performance than the best individual feature set of Ionosphere database. The same thing is observed concerning the poor performance of Adaboost.M1 on Wine and Spectf

Table 2: Summary of fusion performances for all databases

Algorithm / Database	Multiple feature	Wine	Sonar	Ionosphere	Spectf
Best feature set result	90.53	87.55	70.13	93.24	73.37
Adaboost.M1	96.83	-	72.59	94.57	-
Adaboost.RKJ	**97.14**	**90.68**	73.42	-	75.03
Adaboost.RL	97.02	87.57	**73.73**	**94.97**	**75.08**

databases. Adaboost.M1 gives moderate fusion results for the other databases. This shows that the robust combination methods are more appropriate than WMV.

All obtained results and conclusions agree with our hypothesis that WMV is not always appropriate for data fusion applications and that more robust combination methods, as those proposed in this paper, take advantage of the additional information available leading to better classification performance.

5 Conclusion

In this work, we investigate the effectiveness of using robust combination techniques for ensemble methods in the context of data fusion problems. The robust combination rules developed are based on conflict resolution to reduce the influence of outlier models in the ensemble. Fusing multiple sources of data is effective only when appropriate combination method is used. In this context, we have shown that WMV is not always adequate for this kind of problems. However, Adaboost variations namely Adaboost.RL, Adaboost.RKJ, give good fusion performance for most applications. It will be of interest to investigate the effectiveness of employing other robust classifiers combination strategies.

Bibliographie

[1] Freund, Y. and Schapire R.E. (1997) Decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
[2] Briem , G.J. and Benediktsson, J.A. and Sveinsson, J.R. (2002) Multiple Classifiers Applied to Multisource Remote Sensing Data. *IEEE transactions on geoscience and remote sensing*, 40, 2191–2300.
[3] Polikar, R. and Topalis, A. and Green, D. and Kounios, J. and Clark, C.M. (2008) Ensemble based data fusion for early diagnosis of Alzheimer disease. *Information Fusion*, 83–95.
[4] Garcia, M.A. and Puig, D. (2004) Robust Aggregation of Expert Opinions Based on Conflict Analysis and Resolution. *Current Topics in Artificial Intelligence*, 488–497.
[5] Kittler, J.V. (1986) Feature Selection and Extraction. *HPRIP86*, 59–83.

Estimation Non Paramétrique 3

A Robbins-Monro procedure for estimation in semiparametric regression models, *Bernard Bercu and Philippe Fraysse*

Dans un modèle de régression semi-paramétrique, on s'intéresse à l'estimation d'un paramètre de translation et de la fonction de régression. Nous mettons en oeuvre une procédure de Robbins-Monro très efficace et facile à utiliser. Dans un premier temps, on propose un algorithme stochastique récursif de type Robbins-Monro pour l'estimation du paramètre de translation. Cette estimation ne tient pas compte de la fonction de régression. Ensuite, on estime la fonction de régression par un estimateur de type Nadaraya-Watson récursif. Cet estimateur à noyau prend en compte l'estimation préalable du paramètre de translation. Finalement, nous obtenons un algorithme stochastique récursif double permettant à la fois d'estimer le paramètre de translation et la fonction de régression. Pour chacun des estimateurs, on établit la convergence presque sûre et la normalité asymptotique.

Sequential adaptive estimators in nonparametric autoregressive models, *Ouerdia Arkoun*

Ce travail se consacre à l'estimation non paramétrique pour les modèles autorégressifs. Nous considérons le problème de l'estimation d'une fonction inconnue en un point fixe à l'aide de données régies par des modèles autorégressifs. Pour définir le risque associé à l'emploi d'un estimateur et ainsi mesurer la qualité de celui-ci, nous utilisons la fonction de perte liée à l'erreur absolue. Pour un modèle autorégressif non paramétrique où la fonction autorégressive est supposée appartenir à une classe Höldérienne de régularité inconnue, nous obtenons la vitesse de convergence minimax adaptative des estimateurs sur une famille de classes Höldériennes.

Estimation adaptative de la densité de Lévy par une méthode à noyau, *Mélina Bec*

Cet exposé traite de l'estimation non paramétrique et adaptative de la densité de la mesure de Lévy pour des processus de Lévy de sauts purs. Le processus est observé à n instants discrets au pas Delta dans un contexte "haute fréquence". Une collection d'estimateurs à noyau est construite, déduite d'estimateurs pertinents de la fonction caractéristique des accroissements du processus et de sa première dérivée. Deux méthodes de sélection de fenêtre sont présentées et on majore le risque quadratique ponctuel de l'estimateur adaptatif. On illustre ce travail par des simulations basées sur différents exemples.

L'apport du modèle à risques proportionnels de Cox dans la modélisation des prix des actions boursières, *Intissar Mdimagh, Hédi Kortas and Salwa Benammou*

Nous traitons la modélisation des prix des actions boursières à l'aide du modèle semi-paramétrique de Cox avec covariable dépendante du temps. Nous étudions plus particulièrement le phénomène d'apparition du cours boursier minimal. La méthodologie proposée est appliquée sur 79 actions de la bourse de Paris appartenant à l'indice SBF250. Nous proposons un débruitage des données par ondelettes permettant de séparer le signal du bruit sans perte d'information. Nous montrons qu'il ya une amélioration au niveau de la qualité d'ajustement du modèle de Cox.

Estimation dans un modèle de régression semi-paramètrique

Bernard Bercu & Philippe Fraysse

Université Bordeaux 1, Institut de Mathématiques de Bordeaux, UMR CNRS 5251, et INRIA Bordeaux, team ALEA, 351 cours de la libération, 33405 Talence cedex, France.

On se propose d'estimer un paramètre de translation θ et une fonction de régression non paramètrique f dans le modèle de régression défini, pour tout $n \geq 0$, par

$$Y_n = f(X_n - \theta) + \varepsilon_n \tag{1}$$

où (X_n) et (ε_n) sont deux suites indépendantes de variables aléatoires indépendantes et de même loi.

Dans un premier temps, on s'intéresse à l'estimation de θ. Pour ce faire, on met en place un algorithme stochastique récursif de type Robbins-Monro c'est-à-dire :
Supposons qu'on peut trouver une fonction ϕ telle que $\phi(\theta) = 0$. Alors, il est possible d'estimer θ à l'aide de l'algorithme de Robbins-Monro défini par :

$$\widehat{\theta}_0 \in \mathbb{R} \qquad \text{et} \qquad \forall n \geq 0, \quad \widehat{\theta}_{n+1} = \widehat{\theta}_n + \gamma_n T_{n+1} \tag{2}$$

où (γ_n) est une suite de réels positifs qui décroit vers 0 et (T_n) est une suite de variables aléatoires telle que

$$\mathbb{E}[T_{n+1}|\mathcal{F}_n] = \phi(\widehat{\theta}_n)$$

où $\mathcal{F}_n$ est la σ-algèbre engendrée par les événements s'étant produits jusqu'au temps n. Sous des hypothèses classiques sur la fonction ϕ et sur la suite (γ_n), il est connu que $\widehat{\theta}_n$ converge vers θ presque sûrement (cf Duflo, Kushner et Yin). La normalité asymptotique de $\widehat{\theta}_n$ est aussi connue, et on a une loi forte quadratique ainsi qu'une loi du log itéré (cf Hall et Heyde, Mokkadem, Pelletier). Il existe également des versions tronquées de l'algorithme de Robbins-Monro (cf Chen et al, Lelong). L'algorithme adapté à notre problème est très efficace et facile à mettre en place. De plus, il a l'avantage de ne pas prendre en compte la fonction de régression f. On fera cependant attention au fait que dans notre cas on aura la contrainte $|\theta| < 1/4$, et donc on utilisera une version projetée de l'algorithme de Robbins-Monro.

Dans un deuxième temps, on s'intéresse à l'estimation de la fonction de régression f. On propose d'estimer f à l'aide d'un estimateur de type Nadaraya-Watson récursif. Pour l'estimateur classique de Nadaraya-Watson, on sait que celui-ci converge presque sûrement vers la fonction de régression en tout point (cf Noda), et nous avons également une loi du log itéré (cf Hardle et al) ainsi que la normalité asymptotique de l'estimateur (cf Schuster). Dans notre cas, on propose d'utiliser un estimateur de type Nadaraya-Watson récursif de

f qui prend en compte l'estimation préalable du paramètre de translation θ par la suite $\widehat{\theta}_n$. Il est donné, pour tout $x \in \mathbb{R}$, par

$$\forall n \geq 1, \qquad \widehat{f}_n(x) = \frac{\sum_{k=1}^{n} W_k(x) Y_k}{\sum_{k=1}^{n} W_k(x)} \tag{3}$$

avec

$$W_n(x) = \frac{1}{h_n} K\Big(\frac{X_n - \widehat{\theta}_{n-1} - x}{h_n}\Big)$$

où le noyau K est une densité de probabilité et (h_n) est une suite de nombres réels positifs qui décroît vers 0. La principale difficulté ici est de s'affranchir du terme $\widehat{\theta}_{n-1}$ dans le noyau K.

Pour résumer, on obtient un algorithme stochastique double qui permet, en même temps, d'étudier le comportement asymptotique de l'estimateur de Robbins-Monro $\widehat{\theta}_n$ de θ, et de l'estimateur de Nadaraya-Watson $\widehat{f}_n$ de f. Pour chacun des deux estimateurs, on établira la convergence presque sûre et la normalité asymptotique.

Enfin, on illustrera par simulation le comportement de nos deux estimateurs d'une part sur des données simulées et d'autre part sur des données d'électrocardiogramme (ECG).

La première version de ce travail est disponible sur : `http://arxiv.org/pdf/1101.0736`

Abstract

Dans un modèle de régression semi-paramétrique, on s'intéresse à l'estimation d'un paramètre de translation et de la fonction de régression. Nous mettons en oeuvre une procédure de Robbins-Monro très efficace et facile à utiliser. Dans un premier temps, on propose un algorithme stochastique récursif de type Robbins-Monro pour l'estimation du paramètre de translation. Cette estimation ne tient pas compte de la fonction de régression. Ensuite, on estime la fonction de régression par un estimateur de type Nadaraya-Watson récursif. Cet estimateur à noyau prend en compte l'estimation préalable du paramètre de translation. Finalement, nous obtenons un algorithme stochastique récursif double permettant à la fois d'estimer le paramètre de translation et la fonction de régression. Pour chacun des estimateurs, on établit la convergence presque sûre et la normalité asymptotique. Enfin, on illustre le comportement de nos estimateurs via des simulations sur des données simulées et des données réelles (données ECG).

Abstract

This paper is devoted to the parametric estimation of a shift together with the nonparametric estimation of a regression function in a semiparametric regression model. We implement a Robbins-Monro procedure very efficient and easy to handle. On the one hand, we propose a stochastic algorithm similar to that of Robbins-Monro in order to estimate the shift parameter. A preliminary evaluation of the regression function is not necessary for estimating the shift parameter. On the other hand, we make use of a recursive Nadaraya-Watson estimator for the estimation of the regression function. This kernel estimator takes in account the previous estimation of the shift parameter. We establish the almost sure convergence for both Robbins-Monro and Nadaraya-Watson estimators. The asymptotic normality of our estimates is also provided. Finally, we illustrate our semiparametric estimation procedure on simulated and real data (real ECG data).

Mots clés

estimation semi-paramétrique, algorithmes stochastiques, estimation d'un paramètre de translation, estimation d'une fonction de régression, propriétés asymptotiques.

Bibliographie

[1] Bercu B., Portier B. (2008) Kernel density estimation and goodness-of-fit test in adaptive tracking, *SIAM J. Control Optim.*, Vol. 47, No. 5, pp. 2440-2457.
[2] Bigot J., Gadat S. (2010) A deconvolution approach to estimation of a common shape in a shifted curves model, *The Annals of Statistics*, 38 (4), pp. 2422-2464.
[3] Castillo I., Loubes J.M. (2009) Estimation of the law of random shift deformation, *Mathematical Methods of Statistics*, 18, No. 1, pp. 21-42.
[4] Chen H.F., Guo L., Gao A.J. (1988) Convergence and robutness of the Robbins-Monro algorithm truncated at randomly varying bounds, *Stochastic Process, Appl.*, pp. 217-231.
[5] Choi E, Hall P., Rousson V. (2000) Data sharpening methods for bias reduction in nonparametric regression, *The Annals of Statistics*, Vol. 28, No. 5, pp. 1339-1355.
[6] Dalalyan A.S., Golubev K., Tsybakov A.B. (2006) Penalized maximum likelihood and semiparametric second-order efficiency, *The Annals of Statistics*, Vol. 34, No. 1, pp. 169-201.
[7] Dette H., Neumeyer N., Pilz K.F. (2006) A simple nonparametric estimator of a strictly monotone regression function, *Bernoulli*, Vol. 12, No. 3, pp. 469-490.
[8] Devroye L. and Lugosi G. (2001) *Combinatorial Methods in Density Estimation*, Springer Verlag, New York.
[9] Duflo M. (1997) *Random iterative models*, Springer, Berlin.
[10] Gamboa F., Loubes J.M. (2007) Semi-parametric estimation of shifts, *Electronic Journal of Statistics*, Vol. 1, pp. 616-640.
[11] Hall P., Heyde C.C. (1980) *Martingale limit theory and its application*, Academic Press, New York.
[12] Hall P., Huang L-S. (2001) Non parametric kernel regression subject to monotonicity constraints, *The Annals of Statistics*, Vol. 29, No. 3, pp. 624-647.
[13] Hardle W. (1984) A law of the iterated logarithm for nonparametric regression function estimators, *The Annals of Statistics*, Vol. 12, No. 2, pp. 624-635.
[14] Hardle W., Kelly, G. (1987) Nonparametric kernel regression estimation optimal choice of bandwidth, *Statistics*, Vol. 18, No. 1, pp. 21-35.
[15] Hardle W., Tsybakov, A.B. (1988) Robust nonparametric regression with simultaneous scale curve estimation, *The Annals of Statistics*, Vol. 16, No. 1, pp. 120-135.
[16] Hardle W., Janssen, P, Serfling, R. (1988) Strong uniform consistency rates for estimators of conditional functionals, *The Annals of Statistics*, Vol. 16, No. 4, pp. 1428-1449.
[17] Kushner H.J., Clark D.S. (1978) *Stochastic approximation for constrained and unconstrained systems*, Applied Math Science Series, 26, Springer-Verlag, Berlin.
[18] Kushner H.J., Yin G.G. (2003) *Stochastic approximation and recursive algorithms and applications*, Applied Math Science Series, 35, Springer-Verlag, New York, second edition.
[19] Lelong J. (2008) Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions, *Statistics and Probability Letters*, 78, pp. 2632-2636.

[20] Mokkadem A., Pelletier M. (2007) A companion for the Kiefer-Wolfowitz-Blum stochastic approximation algorithm, *The Annals of Mathematical Statistics*, Vol. 35, No. 4, pp. 1749-1772.
[21] Nadaraya E.A. (1964) On estimating regression, *Theory of Probability and its Applications*, Vol. 10, pp. 186-190.
[22] Nadaraya E. A. (1989) *Nonparametric estimation of probability densities and regression Curves*, Kluwer, Dordrecht.
[23] Noda K. (1976) Estimation of a regression function by the parzen kernel-type density estimators, *Annals of the Institute of Statistical Mathematics*, Vol. 28, pp. 221-234.
[24] Parzen E. (1962) On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, Vol. 33, pp. 1065-1076.
[25] Pelletier M. (1998) On the almost sure asymptotic behaviour of stochastic algorithms, *Stochastic processes and their applications*, 78, pp. 217-244.
[26] Robbins H., Monro S. (1951) A stochastic approximation method, *The Annals of Mathematical Statistics*, Vol. 22, No. 3, pp. 400-407.
[27] Robbins H., Siegmund D.. (1971) A convergence theorem for non negative almost supermartingales and some applications, *Optimization methods in Statistics*, Academic Press, pp. 233-257.
[28] Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, Vol. 27, pp. 832-837.
[29] Silverman B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York.
[30] Schuster E.F. (1972) Joint asymptotic distribution of the estimated regression function at a finite number of points, *Annals of Mathematical Statistics*, Vol. 43, No. 1, pp. 84-88.
[31] Stone C.J. (1975) Adaptive Maximum Likelihood Estimators of a Location Parameter, *The Annals of Statistics*, Vol. 3, pp. 267-284.
[32] Tsybakov A. B. (2004) *Introduction à l'estimation non-paramétrique*, Springer-Verlag, Berlin.
[33] Watson G.S. (1964) Smooth regression analysis, *Sankhya*, Vol. 26, pp. 359-372.

Sequential adaptive estimators in nonparametric autoregressive models

Ouerdia ARKOUN

Laboratoire de Mathématiques Raphaël Salem, UMR 6085 CNRS, Université de Rouen, Avenue de l'Université, BP.12, 76801 Saint Etienne du Rouvray (France).
email: Ouerdia.Arkoun@etu.univ-rouen.fr

Abstract

We construct a sequential adaptive procedure for estimating the autoregressive function at a given point in nonparametric autoregression models with Gaussian noise. We make use of the sequential kernel estimators. The optimal adaptive convergence rate is given as well as the upper bound for the minimax risk.

Résumé

Ce travail se consacre à l'estimation non paramétrique pour les modèles autorégressifs. Nous considérons le problème de l'estimation d'une fonction inconnue en un point fixe à l'aide de données régies par des modèles autorégressifs. Pour définir le risque associé à l'emploi d'un estimateur et ainsi mesurer la qualité de celui-ci, nous utilisons la fonction de perte liée à l'erreur absolue. Pour un modèle autorégressif non paramétrique où la fonction autorégressive est supposée appartenir à une classe Höldérienne de régularité inconnue, nous obtenons la vitesse de convergence minimax adaptative des estimateurs sur une famille de classes Höldériennes.

Key words: Adaptive estimation, kernel estimator, minimax, nonparametric autoregression.

1 Introduction

Our problem is the following. Suppose we observe data from the model :

$$y_k = S(x_k)y_{k-1} + \xi_k\,, \quad 1 \le k \le n\,, \tag{1.1}$$

where $x_k = k/n$, y_0 is a constant and $(\xi_k)_{k\in\{1,\dots,n\}}$ are independent and standard Gaussian random variables.

In this paper, similarly to Galtchouk and Pergamenshchikov (2001), we apply the Lepskiĭ procedure to the model (1.1) based on the sequential kernel estimates. We construct the sequential kernel estimator using the method proposed in Borisov and Konev (1977) for the parametric case. It should be noted that to apply the Lepskiĭ procedure the kernel estimators must to have the tail distribution of the Gaussian type. To obtain this property one needs to use the sequential approach. To this end we show some modification of the Levy theorem for discrete time and then, using this result, we show that the sequential kernel estimators have the same form for the tail distribution as a Gaussian random variable. It should be noted that the non sequential kernel estimator does not have the above property in the case of the model (1.1). Thus, in this case, the adaptive pointwise estimation is possible only in the sequential framework.

Let us describe now the sequential kernel estimators. For a constant $H > 0$, we define α_H, such that

$$\sum_{j=1}^{\tau_H-1} Q(u_j)\, y_{j-1}^2 + \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H-1}^2 = H \quad \text{with} \quad u_j = \frac{x_j - z_0}{h}\,,$$

where the kernel $Q(\cdot)$ is the indicator function of the interval $[-1; 1]$, and τ_H is the stopping time defined as follows:

$$\tau_H = \inf\{1 \le k \le n : \sum_{j=1}^{k} Q(u_j)\, y_{j-1}^2 \ge H\}, \tag{1.2}$$

and $\tau_H = n$ when this set is empty.

Note that

$$A_k = \sum_{j=1}^{k} Q(u_j) y_{j-1}^2\,,$$

where h is a positive parameter that we will be define in the next section.

Thus the sequential kernel estimator is written as follows:

$$S_h^*(z_0) = \frac{1}{H}\left(\sum_{j=1}^{\tau_H-1} Q(u_j)\, y_{j-1}\, y_j \,+\, \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H-1}\, y_{\tau_H}\right) \mathbf{1}_{(A_n \ge H)}, \tag{1.3}$$

with $H = nh$. Note that, on the set $\{A_n \ge H\}$ the coefficient $0 \le \alpha_H \le 1$.

Such an estimator is very convenient to estimate the quantity $\mathbf{E}\,|S_h^*(z_0) - S(z_0)|$.

We describe in detail the statement of the problem in section 2 and in Section 3, we illustrate the obtained results by numerical examples.

2 Statement of the problem

The problem is to estimate the function S at a fixed point $z_0 \in]0,1[$, i.e. the value $S(z_0)$. For any estimate $\tilde{S}_n = \tilde{S}_n(z_0)$, the risk is defined on the neighborhood $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ by

$$\mathcal{R}_n(\tilde{S}_n) = \sup_{\beta\in[\beta_*;\beta^*]} \sup_{S\in\mathcal{H}^{(\beta)}(z_0,K,\varepsilon)} N(\beta)\, \mathbf{E}_S |\tilde{S}_n(z_0) - S(z_0)|\,, \tag{2.1}$$

where $N(\beta) = \Big(\dfrac{n}{\ln n}\Big)^{\beta/(2\beta+1)}$ corresponds to the convergence rate of adaptive estimators on class $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ and $\mathbf{E}_S$ is the expectation taken with respect to the distribution $\mathbf{P}_S$ of the vector $(y_1, ..., y_n)$ in (1.1) corresponding to the function S.

We consider model (1.1) where $S \in \mathbf{C}_1([0,1], \mathbb{R})$ is the unknown function. To obtain the stable (uniformly with respect to the function S) model (1.1), we assume that for some fixed $0 < \varepsilon < 1$, the unknown function S belongs to the *stability set*

$$\Gamma_\varepsilon = \{S \in \mathbf{C}_1([0,1], \mathbb{R}) : \|S\| \le 1 - \varepsilon\},$$

where $\|S\| = \sup_{0<x\le1} |S(x)|$. Here $\mathbf{C}_1[0,1]$ is the Banach space of continuously differentiable $[0,1] \to \mathbb{R}$ functions. For fixed constants $K > 0$ and $0 < \beta \le 1$, we define the corresponding *stable local Hölder class* at the point z_0 as

$$\mathcal{H}^{(\beta)}(z_0, K, \varepsilon) = \{S \in \Gamma_\varepsilon \,:\, \Omega^*(z_0, S) \le K\}\,,$$

with

$$\Omega^*(z_0, S) = \sup_{x\in[0,1]} \frac{|S(x) - S(z_0)|}{|x - z_0|^\beta}\,.$$

The regularity $\beta \in [\beta_*; \beta^*]$, is supposed to be unknown, where the interval $[\beta_*; \beta^*]$ is known, $\beta_* > 0$ and $\beta^* \le 1$.

First we give the lower bound for the minimax risk. We show that with the convergence rate $N(\beta)$ the lower bound for the minimax risk is strictly positive.

Theorem 2.1. *The risk* (2.1) *admits the following lower bound:*

$$\liminf_{n\to\infty} \inf_{\tilde{S}_n} \mathcal{R}_n(\tilde{S}_n) \ge \frac{1}{4}\,,$$

where the infimum is taken over all estimators $\tilde{S}_n$.

Now we give the upper bound for the minimax risk of the sequential kernel estimator defined in (1.3). Since β is unknown, one can not use this estimator because the bandwidth h depends on β. That is why we partition the interval $[\beta_*; \beta^*]$ to follow a procedure of Lepskiĭ. Let us set

$$d_n = n/\ln n \quad \text{and} \quad h(\beta) = \left(\frac{1}{d_n}\right)^{\frac{1}{2\beta+1}} .$$

We define the grid on the interval $[\beta_*; \beta^*]$ with the points :

$$\beta_k = \beta_* + \frac{k}{m}(\beta^* - \beta_*), \quad k = 0, \dots, m \quad \text{with} \quad m = [\ln d_n] + 1 .$$

We denote $N_k,\ h_k \ \text{ and } \ \omega(h_j)$ as

$$N_k = N(\beta_k),\ h_k = h(\beta_k) ,$$

and

$$\omega(h_j) = \max_{0\le k\le j} \left(|S^*_{h_j} - S^*_{h_k}| - \frac{\lambda}{N_{k+1}} \right) .$$

We also define the optimal index of the bandwidth as

$$\widehat{k} = \inf \left\{ 0 \le j \le m \,:\, \omega(h_j) \ge \frac{\lambda}{N_j} \right\} - 1 . \tag{2.2}$$

We note that $\omega(h_0) = -\lambda/N_1$ and thus $\widehat{k} \ge 0$. The positive parameter, λ, is chosen as $\lambda > K + e\sqrt{4 + \dfrac{4}{2\beta_* + 1}}$.
The adaptive estimator is now defined as

$$\widehat{S}_n = S^*_{\widehat{h}} \quad \text{with} \quad \widehat{h} = h_{\widehat{k}} . \tag{2.3}$$

The following result gives the upper bound for the minimax risk of the sequential adaptive estimator defined above.

Theorem 2.2. *For all* $0 < \varepsilon < 1$, *we have*

$$\limsup_{n\to\infty} \mathcal{R}_n(\hat{S}_n) < \infty .$$

Remark 2.3. *Theorem 2.1 gives the lower bound for the adaptive risk, i.e. the convergence rate* $N(\beta)$ *is best for the adapted risk. Moreover, by Theorem 2.2 the adaptive estimates* (2.3) *possesses this convergence rate. In this case, this estimates is called optimal in sense of the adaptive risk* (2.1)

Lemma 2.4. *For all $z \geq 2$ and $H > 0$, one has*

$$\mathbf{P}_S(|\zeta_H(h)| > z) \leq 2\, e^{-z^2/8}\,,$$

where

$$\zeta_H = \frac{1}{\sqrt{H}} \left(\sum_{j=1}^{\tau_H - 1} Q(u_j)\, y_{j-1}\, \xi_j + \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H - 1}\, \xi_{\tau_H} \right) \mathbf{1}_{(A_n \geq H)}.$$

3 Numerical simulations

We illustrate the obtained results by the following simulation which is established using Scilab.

The purpose is to estimate, at a given point z_0, the function S defined over $[0;1]$ by $S(x) = |x - z_0|^\beta$. We check that such a function belongs to $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ when $K \geq 1$. The values of z_0 and β are arbitrary, which permit the user to name his choice. As an example, take $z_0 = 1/\sqrt{2}$. Then $\beta_* = 0.6$ is a lower regularity value and $\beta^* = 0.8$ is the higher regularity value.

We simulated n data for the function $S(x) = |x - z_0|^\beta$ for $\beta = 0.7$. We obtained an estimation in constructing the estimator $\hat{S}_n$ defined in (2.3) with the procedure of Lepskiĭ which gives us the optimal bandwidth for the index $\hat{k}$ defined in (2.2).

Numerical results approximate the asymptotic risk of a sequential estimator defined in (2.3) used due to the calculation of an expectation (it performs an average for $M = 15000$ simulations) and the finite number of observations n. Here we calculate for the sequential estimator the quantity $\mathbf{R}_n = \dfrac{1}{M} \displaystyle\sum_{k=1}^{M} |\hat{S}_n^{(k)}(z_0) - S(z_0)|$.

By varying the number of observations n, we obtain different risks listed in the following table:

n	100	1000	5000	10000
$\mathbf{R}_n$	0.284	0.154	0.101	0.087

When taking $\beta = \beta^* = 1$, we obtain

n	100	1000	5000	10000
$\mathbf{R}_n$	0.201	0.097	0.058	0.047

As one can see, the sequential adaptive estimator $\hat{S}_n$ is converging to its true value $S(z_0) = 0$, but this convergence is slow. This is expected since the optimal adaptive convergence rate is $N(\beta) = \left(\frac{n}{\ln n}\right)^{\beta/(2\beta+1)}$. For $\beta = 1$, the results are slightly better than those we got in the first table.

References

[1] Arkoun, O. and Pergamenchtchikov, S. (2008) : Nonparametric Estimation for an Autoregressive Model. Vestnik of Tomsk State University, Ser. *Mathematics and Mechanics* **2** (3), 20 - 30.

[2] Belitser, E. (2000a) : Local minimax pointwise estimation of a multivariate density, *Statisti. Nederletica* **54** (3), 351-365.

[3] Borisov, V.Z. and Konev, V.V. (1977) : Sequential Estimation of Parameters of Discrete Processes, *Automat. and Remote control* **10**, 58-64.

[4] Dahlhaus, R. (1996a) : On the Kullback-Leibler information divergence of locally stationary processes, *Stochastic Process. Appl.* **62** (1), 139–168.

[5] Fourdrinier, D., Konev, V.V. and Pergamenchtchikov, S. (2009) : Truncated Sequential Estimation of the Parameter of a First Order Autoregressive Process with Dependent Noises, *Mathematical Methods of Statistics* **18** (1), 43-58.

[6] Galtchouk, L. and Pergamenshchikov, S. (2001) : Sequential nonparametric adaptive estimation of the drift coefficient in diffusion processes, *Math. Methods Statist.* **10** (3), 316–330.

[7] Helland, I. S. (1981) : Central limit theorems for martingales with discrete or continuous time. *Scet. J. Statist.* **9** (2), 79–94.

[8] Lepskiĭ, O. V. (1990) : A problem of adaptive estimation in Gaussian white noise, *Theory Probab. Appl.* **35** (3), 454-466.

[9] Tsybakov, A. B. (1998) : Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes, *Ann. Statist.* **26** (6), 2420–2469.

Estimation adaptative de la densité de Lévy par une méthode à noyau

Mélina Bec & Claire Lacour

MAP5 UMR CNRS 8145 Université et IUT Paris Descartes, 45 rue des Saints Pères 75006 Paris

&

Laboratoire de Mathématiques Université Paris Sud, 91405 Orsay cedex

Résumé:

Cet exposé traite de l'estimation non paramétrique et adaptative de la densité de la mesure de Lévy pour des processus de Lévy de sauts purs par une méthode à noyau. Le processus est observé à n instants discrets au pas Δ dans un contexte "haute fréquence" ($\Delta = \Delta_n$ tend vers zéro quand $n\Delta_n$ tend vers ∞). Une collection d'estimateurs à noyau est construite, déduite d'estimateurs pertinents de la fonction caractéristique des accroissements du processus et de sa première dérivée. Deux méthodes de sélection de fenêtre sont présentées et on majore le risque quadratique ponctuel de l'estimateur adaptatif. On illustre ce travail par des simulations basées sur différents exemples.

Abstract:

This presentation is concerned with nonparametric adaptive kernel estimation of the Lévy density for pure jump Lévy processes. The process is observed at n discrete instants in "high frequency" context ($\Delta = \Delta_n$ tends to zero while $n\Delta_n$ tends to ∞). We construct a collection of kernel estimators deduced from appropriate estimators of the characteristic function and its first derivative. We present two methods for selecting the bandwidth and we bound the quadratic pointwise risk of the adaptive estimator. We give examples and simulation results for processes fitting in our framework.

Mots clefs. Estimation adaptative; Haute frequence; Processus de Lévy de sauts purs; Estimateurs à noyau non paramétrique.

On considère $(L_t, t \geq 0)$ un processus de Lévy à valeurs réelles dont la fonction caractéristique est donnée par la formule suivante:

$$\psi_t(u) = \mathbb{E}(\exp iuL_t) = \exp\left(t \int_{\mathbb{R}} (e^{iux} - 1)N(dx)\right). \tag{1}$$

On suppose que la mesure de Lévy $N(dx)$ admet une densité $n(x)$ telle que:
(H1) $\int_{\mathbb{R}} |x| n(x) dx < \infty$.

Sous cette hypothèse, $(L_t, t \geq 0)$ est un processus de Lévy de sauts purs. On suppose que les trajectoires sont observées à n instants discrets et on note Δ le pas de l'échantillon. On dispose donc des $(L_{k\Delta})$, $k = 1, ..., n$. L'objectif de ce travail est l'estimation adaptative non paramétrique de la fonction $g(x) = xn(x)$ à partir des observations et lorsque $n \to \infty$. Ce sujet a été récemment traité par de nombreux auteurs. Ces derniers ont pour la plupart développé une procédure d'estimation à partir de la version empirique de la fonction caractéristique $\psi_\Delta(u)$ des accroissements du processus $((Z_k^\Delta = L_{k\Delta} - L_{(k-1)\Delta}, k = 1, \ldots, n))$ et de sa première dérivée. Cette procédure permet de retrouver la densité de Lévy après une transformation de Fourier inverse. Dans le contexte "basse fréquence" (Δ fixé) le problème a été étudié par différents auteurs, on peut citer Jongbloed et Van der Meulen (2006), Watteel et Kulperger (2003) et Comte et Genon-Catalot (2010). Ici, nous nous intéressons au contexte "haute fréquence". En se plaçant dans ce cadre, le probème se simplifie considérablement car $\psi_\Delta(u) \to 1$ lorsque $\Delta \to 0$. Ceci implique que $\psi_\Delta(u)$ n'a pas besoin d'être estimée et sera simplement remplacée par 1 dans la procédure d'estimation. Cette méthode a été mise en place dans Comte et Genon-Catalot (2009) et (2010), partant de l'inégalité:

$$\mathbb{E}\left[Z_k^\Delta e^{iuZ_k^\Delta}\right] = -i\frac{\psi'_\Delta(u)}{\Delta} = g^*(u)\psi_\Delta(u) = g^*(u) + g^*(u)\left(\psi_\Delta(u) - 1\right), \tag{2}$$

où $g^*(u) = \int e^{iux} g(x)dx$ est la transformée de Fourier de g (bien définie sous (H1)). Ensuite, en appliquant une transformée de Fourier inverse à l'estimateur empirique: $\sum_{k=1}^n Z_k^\Delta e^{iuZ_k^\Delta}/(n\Delta)$ avec un paramètre de cutt-off m, un estimateur adaptatif de g est construit par une méthode de sélection de modèle.
Dans notre contexte, on peut appliquer une méthode directe sans passer par l'utilisation de la transformée de Fourier inverse. En effet, on peut déduire de (2) que la distribution empirique:

$$\hat{\mu}_n(dz) = \frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta \delta_{Z_k^\Delta}(dz)$$

converge vers $g(z)dz$. Cela suggère directement l'utilisation d'estimateurs à noyau de la forme:

$$\hat{g}_h(x_0) = K_h \star \hat{\mu}_n(x_0) = \frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta K_h(Z_k^\Delta - x_0)$$

où $K_h(x_0) = 1/hK(x_0/h)$ et $K(-x_0) = K(x_0)$ est un noyau symétrique tel que

$$\int |K(u)|du < \infty \text{ et } \int K(u)du = 1.$$

La formule de l'estimateur $\hat{g}_h(x_0)$ est la suivante:

$$\hat{g}_h(x_0) = \frac{1}{nh\Delta}\sum_{k=1}^n Z_k^\Delta K\left(\frac{Z_k^\Delta - x_0}{h}\right). \tag{3}$$

On étudie ensuite le risque quadratique ponctuel des estimateurs $(\hat{g}_h(x_0))$ et on calcule la vitesse de convergence de ce risque pour $\Delta = \Delta_n \to 0$ et $h = h_n \to 0$ lorsque $n \to \infty$. Le noyau et la fonction g doivent pour cela satisfaire les hypothèses suivante:

- (Ker[1]) K est un noyau symétrique d'ordre $l = \lfloor \beta \rfloor$ et $\int |x|^\beta |K(x)| dx < +\infty$. Le paramètre β sera défini plus loin.
- (Ker[2]) $\|K\|_2 < +\infty$.
- (Ker[3]) $K^* \in \mathbb{L}^1$.
- (G1) $\|g\|_1 < +\infty$.
- (G2) $\|g\|_2 < +\infty$.
- (G3) $g \in H(\beta, L)$.
- (G4) $M := \sup \|g'\|_\infty < +\infty$.
- (G5) $xg \in \mathbb{L}^1$ et $P := \frac{1}{2\pi} \int |g^{*\prime}(u)| du < +\infty$.

On trouve le résultat suivant

Proposition 0.1 *Sous les hypothèses (Ker[1]) à (Ker[3]) , (G1) à (G5) on a*

$$MSE(x_0, h) \leq c_1 h^{2\beta} + c_2 \frac{1}{nh\Delta} + c'_2 \frac{1}{nh} + c'_1 \Delta^2. \tag{4}$$

On pose la condition suivante

$$\Delta^3 \leq \frac{1}{nh}, nh \to +\infty. \tag{5}$$

Alors,

Proposition 0.2 *Sous les hypothèses de la Proposition 0.1 et sous la condition (5) le choix* $h_{opt} \propto ((n\Delta)^{-\frac{1}{2\beta+1}})$ *minimise la borne de risque (4) et donne* $MSE(x_0, h_{opt}) = O((n\Delta)^{-\frac{2\beta}{2\beta+1}})$.

Remarque 0.1 *Ce résultat est un résultat nouveau, les travaux précédents (adaptatifs ou pas) ne traitent pas du risque ponctuel.*

On procède ensuite à la sélection de fenêtre adaptative locale. Pour un $x_0 \in \mathbb{R}$ donné, on définit deux façons de sélectionner la fenêtre $\hat{h}(x_0)$ telle que l'estimateur adaptatif $\hat{g}_{\hat{h}(x_0)}$ atteigne automatiquement la vitesse optimale de convergence (correspondante à

la régularité inconnue de la fonction g). Dans la première méthode, on définit un estimateur $\hat{\beta}(x_0)$ de la régularité inconnue β de g, puis on insère la valeur estimée dans la fenêtre optimale $h_{opt}(\beta)$ déduite de l'étude du risque $\mathbb{L}^2$. La seconde méthode reprend le schéma développé par Goldenschluger et Lepski (2011) pour l'estimation de la densité, en introduisant l'estimateur à noyau suivant:

$$\hat{g}_{h,h'}(x_0) = K_{h'} \star \hat{g}_h(x_0) = K_h \star \hat{g}_{h'}(x_0).$$

On en déduit un choix $\hat{h}_{(x_0)}$ et on définit l'estimateur $\hat{g}_{\hat{h}_{(x_0)}}(x_0)$.

Détails de la première méthode: On considère l'estimateur défini par (3). On pose la condition suivante:

$$\frac{1}{n} << \Delta << \frac{1}{n^{1/3}} \tag{6}$$

Le contexte asymptotique est donc le suivant: $n\Delta \to +\infty$ and $n\Delta^3 \to 0$ et la condition (5) est vérifiée pour tout h.
On définit

$$u_n^2(\beta) = \left(\frac{n\Delta}{\log(n\Delta)}\right)^{-\frac{2\beta}{2\beta+1}} \text{ , } h(\beta) = \left(\frac{n\Delta}{\log(n\Delta)}\right)^{-\frac{1}{2\beta+1}}$$

et on pose $\hat{g}_\beta = \hat{g}_{h(\beta)}$.
Dans la suite, on utilisera la lettre β pour le vrai paramètre. On suppose que β appartient à $A = \{\alpha_1, ..., \alpha_D\}$ où $\alpha_1 < ... < \alpha_D$ et $D = \lfloor (n\Delta)^\xi \rfloor$ avec ξ un réel positif fixé. On choisit $\hat{\beta}(x_0)$ tel que:

$$\hat{\beta}(x_0) = \max\{\alpha \in A,\ \forall \alpha' \in A/\alpha' \leq \alpha,\ |\hat{g}_\alpha(x_0) - \hat{g}_{\alpha'}(x_0)| \leq c u_n(\alpha')\}$$

avec $c \geq c' + 2N + 2B_1$, $c' = 8\|K\|_2\sqrt{P + \|g\|_2^2}\sqrt{2\xi+4}$ et $B_1 = M\|g\|_1\|K\|_1$,et

$$N := \frac{L}{l!}\int |K(v)||v|^\beta dv \text{ et } P := \frac{1}{2\pi}\int |g^{*\prime}(v)|dv. \tag{7}$$

L'estimateur est donc: $\hat{g}_{\hat{\beta}(x_0)}(x_0)$.On a la majoration suivante pour notre estimateur adaptatif:

Théorème 0.1 *Sous les hypothèses de la Proposition 0.1 et si les $|Z_i|$ admettent un moment d'ordre z avec $z > 11 + 4\xi + (3+2\xi)/\alpha_1$, on a sous la condition (6),*

$$\mathbb{E}[|\hat{g}_{\hat{\beta}(x_0)}(x_0) - g(x_0)|] \leq c u_n(\beta) \tag{8}$$

Remarque 0.2 *Le Théorème 0.1 montre que l'estimateur adaptatif atteint automatiquement la vitesse optimale de la Proposition 0.2 à un facteur logarithmique près. Cependant en estimation de la densité adaptative ponctuelle on retrouve également cette perte logarithmique inévitable.*

Détails de la seconde méthode:
On pose:

$$V(h) = C\frac{\log(n\Delta)}{nh\Delta} \text{ avec } C = (c'/2) \times \|K\|_2^2(P + \|g\|_2^2),\ c' \in \mathbb{R}^+, \tag{9}$$

où P est défini dans (7).
On remarque que $V(h)$ a le même ordre que la variance multipliée par $\log(n\Delta)$.
On définit $\hat{g}_{h,h'}(x_0) = K_{h'} \star \hat{g}_h(x_0) = K_h \star \hat{g}_{h'}(x_0)$. On a

$$\hat{g}_{h,h'}(x_0) = \frac{1}{n\Delta}\sum_{k=1}^{n} Z_k^{\Delta} K_{h'} \star K_h(Z_k^{\Delta} - x_0). \tag{10}$$

On pose

$$A(h, x_0) = \sup_{h'} \{|\hat{g}_{h,h'}(x_0) - \hat{g}_{h'}(x_0)|^2 - V(h')\}_+ \tag{11}$$

$h' \in H$, avec $H = \{\frac{j}{M}, 1 \le j \le M\}$, M spécifié plus loin.
On choisit h tel que

$$\hat{h}_{(x_0)} = \arg\min_{h \in H}\{A(h, x_0) + V(h)\}.$$

On note $\hat{h} = \hat{h}_{(x_0)}$.

Théorème 0.2 *Sous les hypothèses de la Proposition 0.1, on suppose qu'il existe β_0 (connu) tel que $\beta > \beta_0$. On définit $H = \{\frac{j}{M}, 1 \le j \le M\}$, avec $M = \left[(n\Delta)^{1/(2\beta_0+1)}\right]$ et on prend c' dans (9) tel que $c' \ge 96(1 \vee \|K\|_\infty)$. Si les Z_i admettent un moment d'ordre z tel que $z \ge 2(3 + 2/\beta_0)$, on a*

$$\mathbb{E}[|g(x_0) - \hat{g}_{\hat{h}}(x_0)|^2] \le C\left\{\inf_{h \in H}\left\{\|g - \mathbb{E}[\hat{g}_h]\|_\infty^2 + V(h)\right\} + \frac{\log(n\Delta)}{n\Delta}\right\}$$

Nous illustrerons notre travail à l'aide de différents exemples et nous présenterons les résultats des simulations. Nous avons implémenté la méthode d'estimation avec différents noyaux. La Figure 1 illustre l'estimation adaptative pour un processus de Lévy gamma ($\alpha = \beta = 1$), K noyau gaussien, n=10000, Δ=0.06. La Figure 2 illustre l'estimation adaptative pour un processus de Lévy Gamma symétrisé ($\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$),$K$ noyau Laplace, n=10000, Δ=0.1.

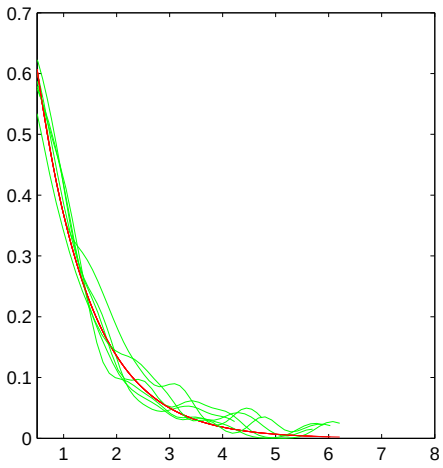


Figure 1:

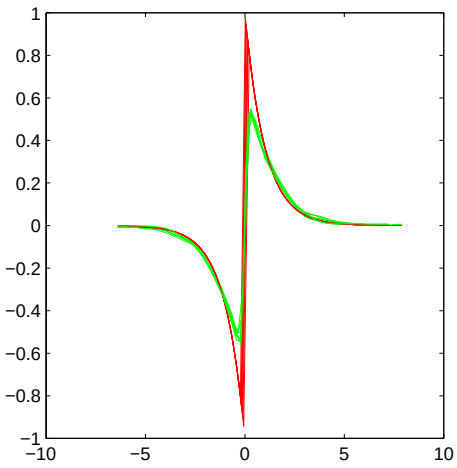


Figure 2:

[1] Comte, F.; Genon-Catalot, V.(2010) Nonparametric adaptative estimation for pure jump Lévy processes.*Annales de l'I.H.P., Probability and Statistics*. 46, no. **3**, 595–617.
[2] Comte, F.; Genon-Catalot, V.(2009) Nonparametric estimation for pure jump Lévy processes based on high frequency data. *Stochastic Process. Appl.* 119 , no. **12**, 4088–4123.
[3] Goldenshluger, A.; Lepski, O.(2010) Bandwidth selection in kernel density estimation: Oracle inequalities and adaptative minimax optimality. Preprint arXiv:1009.1016.
[4] Jongbloed, G.; van der Meulen, F.H.(2006) Parametric estimation for subordinators OU induced processes. *Scand. J. Statist.* 33, 825–847.
[5] Watteel, R.N.; Kulperger, R.J.(2003). Nonparametric estimation of the canonical measure for infinitely divisible distributions. *Journal of Statistical Computation and Simulation*. 73 , no. **7**, 525–542.

L'apport du modèle à risques proportionnels de Cox dans la modélisation des prix des actions boursières

Intissar MDIMAGH & Hedi KORTAS & Salwa BENAMMOU

Intissar MDIMAGH Institut Superieur de Gestion, Université de Sousse, Tunisia. intissar.mdimagh@yahoo.fr
Hedi KORTAS Institut Superieur de Gestion, Université de Sousse, Tunisia. kortashedi@yahoo.fr
Salwa BENAMMOU Faculté de Droit et des Sciences Economiques et Politiques, Université de Sousse, Tunisia. saloua.benammou@fdseps.rnu.tn

Résumé Nous traitons la modélisation des prix des actions boursières à l'aide du modèle semi-paramétrique de Cox avec covariable dépendante du temps. Nous étudions plus particulièrement le phénomène d'apparition du cours boursier minimal. La méthodologie proposée est appliquée sur 79 actions de la bourse de Paris appartenant à l'indice SBF250. Nous proposons un débruitage des données par ondelettes permettant de séparer le signal du bruit sans perte d'information. Nous montrons qu'il ya une amélioration au niveau de la qualité d'ajustement du modèle de Cox.
Mots clés: Modèle de Cox, Durée de survie, Covariable dépendante du temps, Ondelettes.
Abstract We handle the modelling of the prices of the Stock Exchange shares by means of the Cox semi parametric model with time dependent covariate. We study more particularly the phenomenon of appearance of the minimum stock market price. The proposed methodology is applied to 79 shares of the Paris Stock Exchange belonging to the SBF250 index. We propose a denoising of the data by wavelets allowing to separate the signal of the noise without loss of information. We show an improvement of the quality of adjustment of the Cox model.

1 Introduction

Les modèles de durée de vie ou de survie, ont été développés pour des applications en biomédicine, en démographie, en économie, en fiabilité, en actuariat, en finance [6, 9]. L'utilisation de ces modèles pour la modélisation des prix des actions boursières n'a pas été auparavant entreprise.
Dans ce travail, nous nous intéressons à l'utilisation du modèle à risques proportionnels de Cox afin de mieux gérer le phénomène d'apparition du prix minimal des actions.
Malgré les résultats incontestables trouvés par le modèle de Cox, nous avons amélioré la

qualité d'ajustement du modèle à l'aide du débruitage par ondelettes. La méthodologie proposée est appliquée sur des données réelles relative à la bourse de Paris où nous sommes trouvés confrontés à une variable explicative qui évolue dans le temps décrivant le cours de l'action.
Le papier est organisé comme suit : Une présentation du modèle à risques proportionnels de Cox fait l'objet de la section 2. Une brève revue de la théorie des ondelettes est fournie dans la section 3 où la technique de débruitage par ondelettes est décrite. Finalement, dans la section 4, une application du modèle de Cox avec variable dépendante du temps est illustrée sur des données boursières.

2 Le modèle à risques proportionnels de Cox

Le modèle à risques proportionnels est un modèle de régression proposé par Cox en 1972 [3]: c'est le modèle le plus utilisé pour l'analyse des données de survie.
L'analyse des données de survie a pour but de modéliser et d'estimer les lois décrivant la durée de survie T avec ou sans variables explicatives dites covariables.
La durée de survie désigne la variable d'intérêt: il s'agit d'une variable aléatoire positive décrivant la durée passée dans un état donné ou le temps qui s'écoule entre deux événements.
Les fonctions caractérisant la distribution de la durée de survie T possédant une densité de probabilité f et une fonction de répartition F sont essentiellement :

La fonction de risque $h(t)$: mesure le risque instantané de survenue de l'événement d'intérêt sachant qu'il n'est pas encore survécu à l'instant t:
$h(t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < T + \Delta t / T > t]$.

La fonction de survie $S(t)$: mesure la probabilité de ne pas avoir survécu l'événement d'intérêt au moins jusqu'au temps t:
$S(t) = P[T > t]$.

Le modèle de Cox est un modèle de régression semi-paramétrique qui permet de relier la durée de survie à un ensemble de covariables fixes ou dépendantes du temps.
La fonction de survie s'écrit dans le cas d'un modèle de Cox: $S(t/x) = S_0(t)^{exp(x'\beta)}$
où, x est le vecteur de covariables, β correspond au vecteur de paramètres à estimer, $exp(\beta)$ mesure le rapport de risque (RR) qui correspond à l'augmentation du risque instantané de survenue de l'événement à une augmentation de la covariable d'une unité et $S_0(t)$ désigne la fonction de survie de base: où toutes les covariables sont égales à zéro.
Une particularité du modèle à risques proportionnels de Cox est que le rapport des risques instantanés entre deux observations est indépendant du temps.

2.1 L'estimation du modèle de Cox

Les composantes du vecteur β sont estimées à l'aide de la méthode du maximum de vraisemblance partielle de Cox, dont la fonction de vraisemblance est donnée par:
$L(t_{(1)}, ..., t_{(n)}; \beta) = \prod_{i=1}^{n} \frac{exp(x_i'\beta)}{\sum_{k \in R(t_{(i)})} exp(x_k'\beta)}$
où, $(t_{(1)}, ..., t_{(n)})$ sont les $(t_1, ..., t_n)$ classés par ordre croissant, $M(t_{(i)})$ est le nombre d'événements observés à $t_{(i)}$ et $R(t_{(i)})$ est le nombre d'observations à risque à l'instant $t_{(i)}$.

La fonction de survie de base est estimée selon Breslow [2] par :
$\hat{S}_0(t) = \prod_{\{i, t_{(i)} \leq t\}} \left[1 - \frac{M(t_{(i)})}{\sum_{k \in R(t_{(i)})} exp(x_k \hat{\beta})}\right]$

2.2 L'extension du modèle de Cox

Le modèle de Cox peut être étendu en considérant des variables exogènes évoluant au cours du temps. Dans ce cas, x est considéré comme un processus $x(t)_{t \geq 0}$ indexé par le temps et observé que sur l'intervalle de temps [0, T], d'où la non vérification de l'hypothèse des hasards proportionnels [4, 7].
La fonction de vraisemblance partielle de Cox devient dans ce cas:
$L(t_{(1)}, ..., t_{(n)}; \beta) = \prod_{i=1}^{n} \frac{exp(x_i'(t_{(i)})\beta)}{\sum_{k \in R(t_{(i)})} exp(x_k'(t_{(i)})\beta)}$
où $x_i'(t_{(i)})$ désigne le vecteur transposé de covariables à l'instant $t_{(i)}$ pour l'observation i.

3 La théorie des ondelettes

Soit $L^2(\Re)$ l'espace de Hilbert des fonctions d'une variable réelle de carrée intégrable.
Une ondelette est une fonction ψ de $L^2(\Re)$, de moyenne nulle et localisée aussi bien dans le domaine de fréquence que dans le domaine temporelle.
Une famille d'ondelette ψ_{jk} s'obtient en dilatant l'ondelette ψ par un facteur j appelé facteur d'échelle et en la translatant par un facteur k appelé facteur de position.
Les ondelettes sont introduites à l'aide de la notion d'analyse multirésolution (AMR) [10].
On appelle analyse multirésolution de $L^2(\Re)$ la suite des sous espaces fermés V_j de $L^2(\Re)$, $j \in Z$ vérifiant ces 4 conditions suivantes:

1. $\forall j \in Z$, $V_j \subset V_{j+1}$.
2. $\bigcap_{j \in Z} V_j = \{0\}$ et $\bigcup_{j \in Z} V_j = L^2(\Re)$.
3. Pour toute fonction f de $L^2(\Re)$, on a $f \in V_j$ si et seulement si $f(2^{-j}x) \in V_{j+1}$.

4. Il existe une fonction φ dans V_0, dite fonction d'échelle, telle que $\{\varphi(x-k), k \in Z\}$ est une base orthonormée de V_0. Par conséquent, à chaque niveau j, la famille des fonctions $\left\{\varphi_{jk} = 2^{j/2}\varphi(2^j x - k), k \in Z\right\}$ forme une base orthonormée de V_j.

Sous ces conditions, on peut montrer que toute fonction appartenant à $L^2(\Re)$ présente une unique représentation de la forme : $f(x) = \sum_k \alpha_{0k}\varphi_{0k}(x) + \sum_{j=0}^{\infty}\sum_k \beta_{jk}\psi_{jk}(x)$
où α_{0k} et β_{jk} sont respectivement les coefficients d'échelle et d'ondelette.

La technique de débruitage par ondelettes
La technique de débruitage par ondelettes consiste à éliminer par une opération de seuillage des perturbations (bruits) capturées par les coefficients d'ondelettes.
Nous donnons ici l'algorithme de base de débruitage proposé par Donoho [5]:
Etape 1: Décomposition du signal par ondelettes.
Etape 2: Seuillage des coefficients de détails.
Etape 3: Reconstruction du signal par l'application de la transformée inverse par ondelettes.
Donoho présente deux types de seuillage de coefficients d'ondelettes:
seuillage doux (Soft thresholding) et seuillage dur (Hard thresholding).

4 Application à des données boursières

Nous considérons un échantillon de 79 actions cotées à la bourse de Paris appartenant à l'indice SBF250. Nous nous intéressons aux valeurs du cours de ces actions sur les demi-heures d'ouverture et de fermeture de la bourse à la date du 20/03/2009.
Nous définissons la variable à expliquer comme étant le temps en secondes pour l'apparition du cours minimal de l'action durant la demi-heure de fermeture de la bourse.
La variable explicative considérée est formée de l'ensemble des valeurs du cours de l'action durant la demi-heure d'ouverture de la bourse, donc durant 1801 secondes. Nous sommes alors en présence d'une variable indexée par le temps.

4.1 Estimation avant et après débruitage par ondelettes

Nous effectuons dans une première étape une décomposition du cours de l'action par ondelettes. Ensuite, un seuillage doux est appliqué aux coefficients d'ondelettes obtenus. Et enfin, nous reconstruisons le cours débruité. Le seuil de seuillage est fixé selon la procédure Minimax (voir par exemple [8]).
La significativité de la variable prix de l'action est ainsi confirmée ainsi qu'une amélioration de cette significativité suite au débruitage par ondelettes des données comme l'illustrent les tables 1 et 2. En effet, le taux de hasard est donné par 1.712, avec une p-value de 0.0162 et un intervalle de confiance à 95% rangé entre 1.105 et 2.653.

Table 1: Significativité individuelle

	Avant débruitage	Après débruitage
Paramètre estimé	0.37222	0.53765
p-value	0.0326	0.0162
Rapport de Risque(RR)	1.451	1.712
IC (95%) du RR	1.031–2.042	1.105–2.653

Table 2: Significativité globale

Statistique	Avant débruitage	Après débruitage
Rapport de vraisemblance	0.0349	0.0080
Score	0.0354	0.0097
Wald	0.0326	0.0162

Les résultats empiriques nous amènent à conclure que, les valeurs du cours durant la demi-heure d'ouverture de la bourse ont un impact significatif sur la durée ou le temps d'apparition du cours minimal durant la demi-heure de fermeture. Etant donnée que le coefficient associé est positif, nous pouvons dire que plus une action aura des valeurs élevées au début de la journée, plus la durée pour atteindre le prix minimum durant la dernière demi-heure de la journée sera grande.

4.2 Critères de comparaison et sélection du modèle

Pour sélectionner le modèle le plus adéquat parmi les modèles avant et après débruitage par ondelettes, nous allons utiliser trois critères de comparaison: le critère -2 log vraisemblance [4], le critère AIC [1] et le critère BIC [11].
Nous optons donc dans ce travail pour un modèle avec données débruitées puis qu'il donne lieu à la plus petite valeur pour les trois critères comme l'illustre la table 3.

Table 3: Critères de comparaison et sélection du modèle

Critère	Avant débruitage	Après débruitage
-2logvraisemblance	485.899	483.317
AIC	487.899	485.317
BIC	490.203	487.621

5 Conclusion

Dans ce travail nous avons modélisé le phénomène d'apparition du cours minimal de 79 actions du SBF250 en utilisant un modèle de durée semi-paramétrique de Cox avec covariable indexée par le temps. Nous avons proposé ici la technique de débruitage des données par ondelettes nous permettant d'améliorer les résultats.

Bibliographie

[1] Akaike H. (1983) Information measures and model selection, Bulletin of the International Statistical Institue, 50, pp. 277-290.
[2] Breslow N. (1974) Covariance analysis of censored survival data, Biometrics, 30, pp. 89-99.
[3] Cox D R. (1972) Regression models and life-tables, J. Roy Statist. Soc., B 34, pp. 187-220.
[4] David G. Kleinbaum (1996) Survival Analysis: a self-learning text, Springer.
[5] Donoho D. Johnstone I. (1994) Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81, pp. 425-455.
[6] Droesbeke J. J. Fichet B. Tassi P. (1989) Analyse Statistique des durées de vie, Modélisation des données Censurées, ASU, Economica, ed, Paris.
[7] Dupuy J. F. (2002) Modélisation conjointe de données longitudinales et de durées de vie, Thèse. Université René Descartes - Paris 5.
[8] Gencay R., Seluk F. Whitcher B. (2002) An Introduction to Wavelets and Other Filtering Methods in Finance and Economics, Academic Press.
[9] Hougaard P. (2000) Analysis of Multivariate Survival Data, Springer.
[10] MALLAT S. (2000) Une exploration des signaux en ondelettes, Les éditions de l'école polytechnique.
[11] Schwarz G. (1978) Estimating the dimension of a model, The Annals of Statistics 6, pp. 461-464.

Etude de cas Finances

Logistic Models for Credit Scoring, *Waad Bouaguel, Farid Beninel and Ghazi Bel Mufti*

The paper addresses an important problem of inference and learning on different samples in credit scoring case. This growing filed has become a commercially relevant concern with a high practical importance in the banking community. Classical credit systems are often used to evaluate new customer solvency based on previous loan. However, these classical approaches have serious limits and don't take into account the characteristics difference between current customers and the future ones. In order to stay competitive, banking institutions should focus on trying to come up with new and innovative approaches. We show in this case study, using the German credit data set how various forms of linkage can improve the prediction from logistic model to a subpopulation with different characteristics ; otherwise we transfer a classifier from one subpopulation to another. Therefore we obtain seven simple linear mapping models between customers and non customers subpopulations.

La prévision du risque de crédit des banques tunisiennes : Etude comparative entre la régression logistique et la régression logistique à effets aléatoires,*Sami Mestiri and Manel Hamdi*

L'objectif de cet article est de comparer le modèle de la régression logistique versus le modèle de la régression logistique à effets aléatoires 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 les bilans et les comptes de résultats des exercices 1999-2006. Une batterie de 26 ratios ont été calculée à partir de ces documents comptables. En utilisant l'information sur le secteur auquel les entreprises appartiennent, nous avons appliqué le modèle de régression logistique à effets aléatoirs afin de prendre en considération la présence d'une hétérogénéité inobservable. Les résultats obtenus montrent que l'intégration de l'effet sectorielle améliore la qualité des prévisions du modèle en terme de bon classement ainsi que par les résultats obtenus de la courbe ROC.

Vers un modèle intégrateur des antécédants et conséquences du risque perçu par les investisseurs sur le marché boursier tunisien, *Azza Bejaoui and Adel Karaa*

Nous étudions les antécédents et conséquences du risque perçu par les investisseurs individuels présents sur le marché boursier tunisien. En d'autres termes, nous examinons les facteurs influençant le risque perçu par l'investisseur envers le marché boursier ainsi que les réponses comportementales induites par cette perception du risque. A cette fin, un questionnaire a été développé et distribué auprès de 411 investisseurs individuels choisis aléatoirement par les 24 intermédiaires présents sur la Bourse de Tunis. Nous constatons que plus l'investisseur est confiant en soi et optimiste plus sa perception du risque est faible. De même, une perception d'une bonne

qualité de l'information divulguée et une satisfaction vis-à-vis des bénéfices réalloués diminuent le risque perçu envers le marché. Néanmoins, les résultats montrent que l'existence de l'asymétrie d'information augmente, au lieu de réduire, la perception du risque. Par ailleurs, percevoir le risque envers le marché mène à une recherche d'informations intensive (divers types et sources d'information), une bonne performance ainsi qu'une forte intention de réinvestissement. L'utilisation du modèle des équations structurelles nous a permis d'une part de rendre compte de l'importance de la perception du risque dans le processus de prise de décision d'investissement et d'autre part de jeter la lumière sur le rôle de médiateur partiel que joue la performance entre la perception du risque et l'intention de réinvestissement.

Une méthode de traitement des refusés dans le processus d'octroi de crédits, *Asma Guizani, Salwa Ben Ammou and Gilbert Saporta*

L'analyse du risque de défaillance de l'emprunteur a toujours constitué, pour une banque, le cœur de la problématique de l'analyse financière qui accompagne toute demande de crédit. Disposer de modèles statistiques de prévision de défaillance est donc devenue indispensable pour une banque. Ces techniques de prévision ont pour objectif d'évaluer la solvabilité future des clients potentiels et cela en calculant le score d'octroi de crédit : c'est le principe du crédit scoring. Ces modèles de crédit scoring sont basés sur l'historique de remboursement de crédit par les clients admis par l'organisme de crédit, l'information sur les refusés manque donc systématiquement. Il en résulte que les modèles sont construits sur des échantillons non représentatifs de la population totale, ce qui aboutit à des résultats biaisés. Pour remédier à ce problème, on a eu recours au concept de traitement des refusés qui a pour but de corriger ce biais en réintégrant dans l'échantillon les dossiers refusés. De nombreuses méthodes de traitement des refusés existent dans la littérature. Parmi elles on peut citer la technique d'augmentation dite aussi de "'re-pondération"' ou "'re-weighting"', la technique d'extrapolation, la technique de reclassification, la technique de Parcelling et le groupe de contrôle. La méthode d'inférence de rejet que nous utilisons dans notre cas est celle de l'augmentation simple et les techniques statistiques que nous adoptons pour la construction des modèles de score sont la régression PLS et l'analyse factorielle discriminante et ceux pour leur robustesse et la simplicité de l'interprétation des résultats.

Logistic Models for Credit Scoring

Bouaguel Waad[1], Beninel Farid[2]& Bel Mufti Ghazi[3]

[1]*LARODEC,ISGT, University of Tunis,bouaguelwaad@mailpost.tn.*
[2]*CREST-ENSAI & UMR 6086, Campus de Ker Lann, France, fbeninel@ensai.fr.*
[3]*ESSEC, University of Tunis, belmufti@yahoo.com.*

Abstract

The paper addresses an important problem of inference and learning on different samples in credit scoring case. This growing filed has become a commercially relevant concern with a high practical importance in the banking community. Classical credit systems are often used to evaluate new customer solvency based on previous loan. However, these classical approaches have serious limits and don't take into account the characteristics difference between current customers and the future ones. In order to stay competitive, banking institutions should focus on trying to come up with new and innovative approaches. We show in this case study, using the German credit data set how various forms of linkage can improve the prediction from logistic model to a subpopulation with different characteristics; otherwise we transfer a classifier from one subpopulation to another. Therefore we obtain seven simple linear mapping models between customers and non customers subpopulations.

keywords: *Learning and classification, Models for finances, Logistic model, Subpopulation links, Subpopulations mixture.*

Résumé

Le papier aborde un problème important d'inférence et d'apprentissage sur des échantillons différents dans le contexte du credit scoring, ce domaine en croissance est devenu l'une des plus hautes préoccupations dans la communauté bancaire. Les systèmes de crédit classiques sont souvent utilisés pour évaluer la solvabilité d'un nouvel emprunteur en se basant sur son historique de prêt. Cependant, ces approches classiques présentent des sérieuses limites et ne prennent pas en compte la différence de caractéristiques entre les clients actuels et les futurs clients. Pour rester compétitives, les institutions bancaires doivent essayer d'inventer des approches innovatrices. Nous montrons dans cette étude de cas en utilisant un jeu de données de crédits à la consommation d'une banque allemande comment la présence de divers forme de liens entre des sous populations améliore la qualité de prédiction du modèle logistique d'une sous-population avec des caractéristiques différentes. Nous obtenons par la suite sept modèles logistique de liens entre la sous-population des clients actuels et celle des clients futurs.

Mots-clés: *Apprentissage et classification, Modèles pour les finances, Model logistique , liens de sous-populations, Sous-populations mixtes.*

1 Introduction

Credit scoring is used to assign credit applicants to two classes of risk: good and bad. Over the time, credit scores have become a fundamental factor in the prediction of credit applicants' behavior. The credit score of the new applicant is a number that summarize his characteristics in order to be analyzed and modeled through scorecard (i.e. credit models). The literature on credit scoring and credit scorecard models is quite vast, authors like Hand and Henley (1997) and Thomas et al. (2002) have been interested to define and discuss credit scoring concepts. Many techniques have been suggested to develop credit scorecard including traditional statistical , which involve discriminant analysis (Fisher, 1936), logistic regression (Henley and Hand, 1997) ...

The used data to construct a scorecard is generally obtained from a sample of applicants to whom credit has already been granted, and for whom it is known whether or not the creditor was reliable. If we consider that a new applicant is classified according to his historical loan, what about applicant who have no credit history? Shall we classify him according to a practiced classification rule or try to get new specific rules.

In this case study applicants with historical loans are represented through the customers' borrower subpopulation and those without historical loans will be represented through the non customers' borrower subpopulation. Borrower's behavior is described by a binary target variable denoted Y, value taken by this last one supplies a basic element in credits' granting decision, $Y = 0$ when the borrower presents problem and $Y = 1$ otherwise. Beside this variable, every borrower is also described by a set of description variables $(X_1, X_2, \ldots, X_d)$ informing about the borrower and about his accounts' functioning. The sample of loans' applicants results from a heterogeneous population formed by the previous two subpopulations.

We will focus in non customers' subpopulation credit worthiness evaluation. Assuming that sample size of this subpopulation is small, we will adopt one of the most used and efficient traditional statistical technique, which is logistic regression. However, using a learning sample from the non customers' subpopulation to build their logistic model isn't appropriate because of the subpopulation small size. to resolve this size problem we use besides non customers' sample a design sample, drown from another population considered slightly different (e.g. customers' subpopulation). In a multinomial context, Biernacki et al. (2002) proved that two slightly different populations are linked through linear relations. Estimation of nonlabeled sample allocation rules was obtained via estimating the linear relationship parameters, using constraints models on the linear relationships.

This approach proved to be efficient in biological context and many extension of this paper was proposed, including Bouveyron and Jacques (2009) as well as Beninel and Biernacki (2009), given previous results we will present seven logistic mapping models between customer and non customers' subpopulations. All logistic models will be compared and the best ones will be used for non customers' subpopulation.

2 Logistic regression models

Logistic regression is one of the most common techniques for credit scoring (Hand and Henley, 1997). Many authors as Fan and Wang (1998) recommend it because of its simplicity and explainability. Logistic regression model supplies a linear function of descriptors as discrimination tool. Model form is given by

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta^T \mathbf{x}_i, \tag{1}$$

where p_i is the posteriori probability, defined as the probability that an individual i have the modality 1 for given values taken by descriptors, $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^d)$ is the vector of observed value taking by description variables. We denote by $\beta^T = (\beta_1, \beta_2, \ldots, \beta_d)$ the vector of variables effect and β_0 the intercept.
Computing the posteriori probability p_i, allows to assign each borrower to his appurtenance group $\{Y_i = 1\}$ if $p_i \geq 0.5$ and $\{Y_i = 0\}$ otherwise. Let us remind that we deal with the problem of discrimination in case of subpopulations' mixture, where the two subpopulations of interests are the subpopulation of borrowers customers and the subpopulation of borrowers non customers, denoted respectively Ω and Ω^*, for which we associate the two following posteriori probability p and p^*.

Given two learning samples $S_L = \{(\mathbf{x}_i, Y_i) : i = 1, ..., n\}$ and $S_L^* = \{(\mathbf{x}_i^*, Y_i^*) : i = 1, ..., n^*\}$ drown from Ω and Ω^*, where the couples $(\mathbf{x}_i, Y_i)$ and $(\mathbf{x}_i^*, Y_i^*)$ are independent and identically distributed (i.i.d.) realizations of the random couples $(\mathbf{x}, Y)$ and $(\mathbf{x}^*, Y^*)$. We consider the logistic model over Ω, as given by

$$F(\mathbf{x}_i, \theta) \quad = p_i \quad = P(Y_i = 1|\mathbf{x}_i, \theta) = \frac{exp^{\beta_0 + \beta^T \mathbf{x}_i}}{1 + exp^{\beta_0 + \beta^T \mathbf{x}_i}}, \tag{2}$$

and over Ω^*

$$F^*(\mathbf{x}_i^*, \theta^*) \quad = p_i^* \quad = P(Y_i^* = 1|\mathbf{x}_i^*, \theta^*) = \frac{exp^{\beta_0^* + \beta^{*T} \mathbf{x}_i^*}}{1 + exp^{\beta_0^* + \beta^{*T} \mathbf{x}_i^*}}, \tag{3}$$

where $\theta = \{(\beta_0 || \beta^T) \in \mathbb{R}^{d+1}\}$ and $\theta^* = \{(\beta_0^* || \beta^{*T}) \in \mathbb{R}^{d+1}\}$ are the sets of all parameters to be estimated respectively over Ω and Ω^*. We denote by $(\beta_0 || \beta^T)$ and $(\beta_0^* || \beta^{*T})$ the concatenations of the intercept and the vector of variables effect over Ω and Ω^*.

In our case we assume that an experienced rule, to predict on the subpopulation Ω is known and we have a small learning sample from the subpopulations Ω^*. From available data we want to get a new allocation rule over Ω^*. According to Beninel and Biernacki (2009) and Bouveyron and Jacques (2009) links between subpopulations could exist and consequently information on Ω could provide some information on Ω^*. Existence of particular connections between the variables distributions lead to relations between the parameters of their respective logistic regression models given by (2) and (3).

In this context Beninel and Biernacki (2009) supposed the existence of an application $\phi_k : \mathbb{R}^d \to \mathbb{R}^d$ linking in law the random variables vectors of Ω and Ω^*. Then

$$\begin{array}{l} \mathbf{x}^*_{|Y^*=k} \sim \phi_k(\mathbf{x}_{|Y=k}) = [\phi_{k1}(\mathbf{x}_{|Y=k}), ..., \phi_{kd}(\mathbf{x}_{|Y=k})]^T \\ \mathbf{x}^*_{|Y^*=k} \sim \Lambda_k \mathbf{x}_{|Y=k} + \alpha_k, \end{array} \tag{4}$$

where Λ_k is a diagonal matrix defined over $\mathbb{R}^{d\times d}$ and α_k is a vector of $\mathbb{R}^d$. The two previous expressions show the relations between the two subpopulations' variables distributions, which provide certain links between the parameters of the two subpopulations scoring functions. Anderson (1982) proved the existence of a link between the parameters of the mixture Gaussian model and those of corresponding logistic model. Links between the two subpopulations can be obtained in a stochastic case where, the variables vectors $\mathbf{x}$ and $\mathbf{x}^*$ defined over Ω and Ω^* are Gaussian, homoscedastic conditionally in the groups. We obtain the following expressions for β_0^* and β^*:

$$\beta_0^* = \alpha + \beta_0 \text{ and } \beta^* = \Lambda\beta. \tag{5}$$

Consequently, the scoring function obtained by replacing the parameters β_0^* and β^* in equation (3) is given by:

$$F^*(\mathbf{x}_i^*, \theta, \varrho) = \frac{exp^{\beta_0+\alpha+(\Lambda\beta)^T\mathbf{x}_i^*}}{1 + exp^{\beta_0+\alpha+(\Lambda\beta)^T\mathbf{x}_i^*}}, \tag{6}$$

here $\varrho = \{(\alpha||\Lambda) \in \mathbb{R}^{d+1}\}$ is the set of transition parameters to be estimated, where $(\alpha||\Lambda)$ is the concatenation of the scalar α and the diagonal of the matrix Λ. Estimation of links between Ω and Ω^* subpopulations is done through several logistic intermediary sample models of connections. These models are listed and summarized in Table 1.

3 Empirical analysis

To evaluate our models, we choose to use the German credit data (Fahrmeir and Tutz, 1994). The data set cover a sample of 1000 credit consumers where 700 instances are creditworthy applicants and 300 are not. Each applicant is described by a binary target variable $Kredit$, $Kredit = 1$ for creditworthy and $Kredit = 0$ otherwise. 20 other input variables are assumed to influence this target variable, we use the variable $Laufkont$ (balance of current account) to separate the data set in two subpopulations: when $Laufkont > 1$ we obtain the customers subpopulation of size 726 and when $Laufkont = 1$ we obtain the non customers subpopulation of size 274. Afterward, we conversed the subpopulation of borrowers non customers in two samples: a learning sample S_L^* and a test sample S_T^*.

To obtain a robust estimate of our seven models performance, our simulations involves taking 50 random design of size $n \in \{50, 100, 150, 200\}$ and test sample splits from the non customers subpopulation. Test error rate, Type II and Type I error are used as performance measures.

Table 1: *Links models*

Models	**Learning sample**	**Parameters**	**Descreptions**
$M1$	Ω	$\alpha = 0 \quad \Lambda = I_d$	The score functions are invariable.
$M2$	S_L^*	$\alpha = 0 \quad \Lambda = \lambda I_d$	The score functions of the two subpopulations differ only through the scalar parameter λ.
$M3$		$\alpha \in \mathbb{R} \quad \Lambda = I_d$	The score functions of the two subpopulations differ only through the scalar parameter β_0^*.
$M4$		$\alpha \in \mathbb{R} \quad \Lambda = \lambda I_d$	The score function of the two subpopulations differ through the couple (β_0^*, λ).
$M5$		$\alpha = 0 \quad \Lambda \in \mathbb{R}^{d\times d}$	The score functions of the two subpopulations differ only through the vectoriel parameter β^*.
$M6$		$\alpha \in \mathbb{R} \quad \Lambda \in \mathbb{R}^{d\times d}$	There is no more stochastic link between the logistic discriminations of the two subpopulations. All parameters are free.
$M7$	$S_L^* \cup \Omega$	$\beta_0 \cup \alpha \in \mathbb{R} \quad \beta \cup \Lambda \in \mathbb{R}^{d\times d}$	The model consist in introducing as observations, all the borrowers (customer and non customers) and to apply a simple logistic regression. This consists in the joined estimation of Ω parameters and transition parameters .

Table 2 shows that the most banks' practiced model $M1$ seem to be the least successful model once applied to non customers borrowers' data. Models $M1$ and $M7$ have the most raised Type I error rate, It seems that these two models have greater difficulty in predicting non-reliable clients than reliable ones. This confirms the difference between the two studied subpopulations.

From the Table 2 it's obvious that models $M5$ and $M6$ possesses the most raised rate of Type II error. These models are considered careful but they are less efficient in the reliable applicants prediction. These two models have low rates of Type I error, they might be considered as good classifier. However, models $M3$ and $M4$ seems to be the more

Table 2: *Results summary (learning sample size* $= 200$*)*

	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
Test error	0.367	0.337	0.308	0.308	0.321	0.315	0.334
Type II error	0.185	0.283	0.294	0.296	0.305	0.304	0.203
Type I error	0.336	0.271	0.220	0.218	0.250	0.273	0.278

suitable models for the prediction of non customers behavior to pay back loan. Models $M3$ and $M4$ achieved almost the same and the lowest rate of misclassified instances and almost the same Type II and the lowest Type I error rates. These last ones are the best predictive models because their constant is calculated from the non customers learning sample independently of customers sample, what confirm the existence of a certain link between the two subpopulation Ω and Ω^*.Simulations with others sample size confirms the summarized results in Table 2.

4 Conclusion

The credit worthiness problem was considered in this paper, we studied a population of insufficient size and seven logistic models were discussed. A comparison of the seven models performance was done and the models $M3$ and $M4$ was selected as the best classification model for the non customers subpopulation. We envisage as perspective, to apply logistic regression using non-linear links between the two subpopulations. We also can apply a fusion approach for the seven logistic model results.

References

[1] Anderson, J.A. (1982) Logistic discrimination. In Handbook of Statistics, 2:169-191.

[2] Beninel, F. and Biernacki, C. (2009) Updating a logistic discriminant rule: Comparing some logistic submodels in credit-scoring. *In International Conference on Agents and Artificial Intelligence*, pages 267-274, France.

[3] Biernacki, C., Beninel, F. and Bretagnolle, V. (2002) A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58:387-397.

[4] Bouveyron, C. and Jacques, J. (2009) Modèles adaptatifs pour les mélanges de régressions. *In 41èmes Journées de Statistique, SFdS*, Bordeaux, France. inria-00386638, version 1-22.

[5] Fahrmeir, L. and Tutz, G. (1994) Multivariate Statistical Modelling Based on Generalized Linear Models (Springer Series in Statistics). *Springer*, 2nd edition, April.

[6] Fan, X. and Wang, L. (1998) Comparing linear discriminant function with logistic regression for the two-group classification problem. *In Annual Meeting of American Educational Research association*, pages 265-286.

[7] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-188.

[8] Hand, D.J. and Henley, W.E. (1997) Statistical classification methods in consumer credit scoring: a review. *Journal Of The Royal Statistical Society Series A*, 160(3):523-541.

[9] Thomas, L.C., Crook, J. and Edelman, D. (2002) Credit Scoring and Its Applications. *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA.

La prévision du risque de crédit des banques tunisiennes : Etude comparative entre la régression logistique et la régression logistique à effets aléatoires

Sami Mestiri

Unité de recherche EAS-Mahdia
Faculté des sciences économiques et de gestion de Mahdia,
Université de Monastir, Tunisie.

Manel Hamdi

Unité de recherche IFGT-Tunisia
Faculté des sciences économiques et de gestion de Tunis,
Université El Manar , Tunisie.

Résumé : L'objectif de cet article est de comparer le modèle de la régression logistique versus le modèle de la régression logistique à effets aléatoires 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 les bilans et les comptes de résultats des exercices 1999-2006. Une batterie de 26 ratios ont été calculée à partir de ces documents comptables. En utilisant l'information sur le secteur auquel les entreprises appartiennent, nous avons appliqué le modèle de régression logistique à effets aléatoires afin de prendre en considération la présence d'une hétérogénéité inobservable. Les résultats obtenus montrent que l'intégration de l'effet sectorielle améliore la qualité des prévisions du modèle en terme de bon classement ainsi que par les résultats obtenus de la courbe ROC.
Mots clés : Prévision, risque de crédit, régression logistique, régression logistique à effets aléatoires, courbe ROC.

Abstract : The aim of this paper is to compare the model of logistic regression versus logistic regression model with random effects in order to predict the credit risk of banks in Tunisia. The sample includes 528 Tunisian firms from different sectors of activities that we have balance sheets and income statements for fiscal years 1999-2006. A battery of 26 ratios were calculated from these records. Using information on the sector to which firms belong, we applied the logistic regression model with random effects to take into account the presence of unobserved heterogeneity. The results obtained show that the integration of sectoral effect improves the quality of model predictions in terms of good classification as well as by the ROC curve result.
Key words : Forecasting ; Logistic mixed distress scoring ; Curve ROC.

Introduction et sommaire

Les banques et les organismes financiers se trouvent face à l'obligation de parvenir à une meilleure gestion du risque de crédit en développant des outils statistiques dans le but de prévoir la détresse financière des entreprises. Dans ce contexte, et poursuivant le travail pilote de Fisher (1933) sur l'analyse discriminante, nombreuses recherches ont été menées. Durand (1941) fut le premier à reconnaître la possibilité d'utiliser des techniques statistiques pour discriminer entre bons et mauvais emprunteurs. En effet, le crédit scoring est le plus utilisé par les institutions financières, notamment, dans le domaine du crédit à la consommation et du crédit aux professionnels (Thomas, 2006) et aux firmes de petite taille (Ono, 2006) et aux petites et moyennes entreprises (Dietsh, 2003). Les premiers modèles de scores remontent aux travaux pionniers de Beaver (1966) et Altman (1968). Ainsi, plusieurs travaux ont été effectués, qui permettent d'associer une probabilité de défaillance à un score observé, notamment ceux de Conan et Holder (1984) et de Bardos (1984 et 1991). Même en France, la centrale de bilan de la banque de France a fortement contribué à la diffusion des modèles de scoring (Bardos, 2000).

Face à la contrainte de la normalité de l'analyse discriminante, certain chercheurs ont préféré utiliser des autres techniques telles que le modèle de la régression logistique qui sera présenté dans ce papier. En effet, Press et Wilson (1978) ont utilisé des données de ratios en coupe transversale pour examiner si les coefficients de la fonction de score estimés a partir du modèle de régression logistique sont des déterminants valides de la faillite des entreprises. Cependant, des informations importantes pourraient être omises en utilisant seulement une analyse en coupe transversale. L'analyse de données longitudinale est une technique appropriée pour traiter ce genre de problème, parce qu'elle tient compte des propriétés des effets non observables qui peuvent être dûs aux regroupements de l'échantillon étudié en classe. Dans ce cas, la modélisation des effets peut intervenir dans l'explication du phénomène étudié. La partie explicative du modèle est raffinée par la combinaison linéaire de ces deux types d'effets : les effets fixés et les effets aléatoires.

En ce concerne le modèle basé sur la régression logistique, nous avons sélectionné 8 ratios significatifs parmi 26 ratios de l'étude. Étant donné la structure longitudinale des données de notre étude, nous avons appliqué le modèle de la régression logistique à effets aléatoires dans le cadre de calcul du risque de la détresse en prenant en considération la présence d'une source d'hétérogénéité individuelle. Le modèle de la régression logistique à effets aléatoires s'écrit sous la forme suivante :

$$\begin{aligned} log\left(\frac{P_{ij}}{1-P_{ij}}\right) &= \beta_1 R_{7,ij} + \beta_2 R_{9,ij} + \beta_3 R_{10,ij} + \beta_4 R_{14,ij} \\ &\quad + \beta_5 R_{20,ij} + \beta_6 R_{21,ij} + \beta_7 R_{23,ij} + b_i, \end{aligned} \tag{1}$$

tel que P_{ij}=$P(Y=1|R_{ij})$ avec $i=1,...,20$ et $j=1,...,n_i$ est la probabilité a posteriori d'appartenance au groupe d'entreprises en détresse, R_{ij} sont les variables explicatives ratios financières et b_i est l'effet spécifique sectoriel qu'on suppose suivre la loi Normale. Ainsi, nous avons associé aux ratios un effet spécifique sectoriel pour modéliser l'hétérogénéité des entreprises par un effet sectoriel.

Le modèle de régression logistique à effets aléatoires (1) a été estimé par la méthode du maximum de vraisemblance marginale (Breslow et Clayton (1993)) . Nous avons utilisé le package (glmmPQL) du logiciel R pour l'estimation des paramètres du modèle de régression logistique à effets aléatoires. La table 1 rapporte les résultats d'estimation du modèle (1) pour les données de notre échantillon :

	Val. estimés	*Pouv. discrim.*	t value	*p-value*
(constante)	-2.258303		-18.03	0.0000
R_7 : Rotation de l'actif	0.235746	0.0016	3.87	0.0001
R_9 : Rentabilité économique	8.742052	0.5414	8.36	0.0000
R_{10} : Rentabilité des capitaux	-10.65694	0.4506	-8.40	0.0000
R_{14} : Taux de rentabilité des capitaux	0.033662	0.0000	1.79	0.0740
R_{15} : Rotation des capitaux	-0.002738	*0.0000*	-1.65	0.0993
R_{20} : Couverture des immobilisations	0.237643	0.0062	-4.82	0.0000
R_{21} : Capacité d'endettement	-0.238740	0.0000	-2.64	0.0084
R_{23} : Ratio de charges financières	-0.272702	0.0000	-2.61	0.0091

TAB. 1 – Les coefficients estimés du modèle de régression logistique à effets aléatoires

Le pouvoir discriminant du ratio R_k est défini par le rapport : $\frac{\sigma_k^2\beta_k^2}{\sum \sigma_k^2\beta_k^2}$ avec σ_k est l'écart type du ratio R_k. Il exprime l'influence du ratio dans la fonction de score. D'après la table 1, les ratios R_9 et R_{10} jouent un rôle capital dans la formation de la fonction de score des entreprises puisque ce ratio a un pouvoir discriminant de l'ordre de 99%. En adoptant la spécification du modèle (1), nous remarquons que l'effet estimé de la variable R_9 (la rentabilité économique) a un signe positif. Comme le ratio la rentabilité économique est égale au rapport des frais financier sur l'actif total. Cela signifie que l'augmentation des frais financiers fait diminuer la rentabilité économique ce qui explique l'accroissement de la probabilité d'être en détresse. Par contre la variable R_{10} (la rentabilité des capitaux investis) qui est égale au rapport du résultat net sur l'actif total présente un signe négatif ce qui induit que l'augmentation des résultats net implique une diminution de risque de défaillance.

Après l'intégration de l'effet sectoriel dans le modèle de régression logistique, nous avons abouti aux estimations présentés dans la table la table (2). Ces estimations des effets aléatoires sectoriels présentent un classement des secteurs de moins risqués aux plus risqués. Autrement dit d'après les résultats de la table (2), le secteur " Commerce,

codes	Les secteurs	Effets aléatoires
1	Commerce, réparations automobile et d'articles domestiques	-4,401
2	Métallurgie et travail des métaux	-2,943
3	Industrie du caoutchouc et des plastiques	-1,480
4	Industrie du cuir et de la chaussure	-1,009
5	Agriculture chasse sylviculture	-0,768
6	Fabrication de machines et équipements	-0,654
7	Santé et action sociale	-0,596
8	Industries agricoles et alimentaires	-0,334
9	Immobilier locations et services aux entreprises	-0,256
10	Fabrication d'autres produits minéraux non métalliques	0,211
11	Industrie textile et habillement	0,284
12	Industrie chimique	0,377
13	Transports et communications	0,473
14	Fabrication équipements électriques et électroniques	0,551
15	Extraction de produits non énergétiques	0,584
16	Industrie du papier et du carton édition et imprimerie	0,597
17	Construction	0,860
18	Hôtels et restaurants	1,045
19	Industries agricoles et alimentaires	1,198
20	Autres industries manufacturières	6,261

TAB. 2 – Les coefficients estimés des effets aléatoires

réparations automobile et d'articles domestiques " est le secteur le moins risqué, puisqu'il admet -4.401 comme effet aléatoire. Par contre nous avons enregistré un effet de 6.261 pour le secteur "Autres industries manufacturières" que nous pouvons considérer comme le secteur le plus risqué.

Après avoir déterminé des fonctions de score de la détresse, il faut en évaluer leurs efficacité. 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 Characteristic" en calculant les indices associés tels que l'aire sous la courbe de ROC.

La table 3 présente les taux d'erreur de classement qui est égale au nombre de mauvais classement rapporté à l'effectif total. Le taux d'erreur de classement égale à 14% pour le modèle de régression logistique classique et 11.9% pour le modèle de régression logistique à effets aléatoires c.à.d une amélioration de prédiction de 3.1%. Ce qui prouve l'importance de l'intégration des effets sectoriels dans le calcul de risque de la détresse.

	La régre. logistique classique			La régre logistique à effets aléatoires		
	$\hat{Y}=1$	$\hat{Y}=0$	Total	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Y=1$	5	2	7	20	4	24
$Y=0$	84	522	606	69	520	589
Le taux d'erreur	0.140			0.119		

TAB. 3 – Matrice de confusion des modèles estimés pour l'échantillon test

De même dans le but de comparer le modèle de régression logistique classique et le modèle de régression logistique à effets aléatoires, nous présentons la courbe ROC de chaque modèle. Ce courbe est un outil graphique qui permet d'évaluer et de comparer globalement le comportement des fonctions de scores (Pepe(2000). D'après la courbe ROC, il est évident que la règle de classification basée sur la régression logistique à effets aléatoires est plus performante que celle basée sur la régression logistique standard. Ceci nous amène à conclure que la fonction de score issue du modèle de régression logistique à effets aléatoires est meilleur que celle obtenue à partir du modèle de régression logistique standard.

A partir de la courbe ROC, nous pouvons synthétiser un indicateur qui reflète le pouvoir prédictif du modèle. En fait, l'aire sous la courbe ROC (AUC) mesure la qualité de discrimination du modèle et traduit la probabilité qu'une entreprise saine présente un score supérieur au score d'une entreprise en détresse. L'AUC du modèle de régression logistique égale à 0.684 par contre 0.811 pour le modèle de régression logistique à effets aléatoires. Ces valeurs obtenues plus sont proches de un. Ce qui montre l'avantage de l'intégration de l'effet sectoriel et son impact sur le pouvoir prédictif du modèle de régression logistique.

Bibliographie

[1] Altman, E. I. (1968). "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *The Journal of Finance*, 23 :589–609.

[2] Bardos, M. and Zhu, W. H. (1997). "Comparaison de l'analyse discriminante linàaire et des réseaux de neurones. application à la détection de défaillance d'entreprises." *Revue Statistique Appliquée.*

[3] Beaver, W. H. (1966). "Financial ratios as predictors of failure". *Journal of Accounting Research*, 4 :71–111.

[4] Breslow, N. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models". *Journal of the American Statistical Association*,88 :9 – 25.

[5] Bardos, M. (1991). "Liaison entre le risque et la taille ; influence du risque et de la taille sur le taux d'intérêt" *Cahiers Economiques et Monétaires*, n° 38, pp. 49-104.

[6] Efron, B., (1975) : "The efficiency of logistic regression compared to normal discriminant analysis" *Journal American Statistical Society* n°70, pp. 892-898.

[7] Brostr, A. (2003). "Generalized linear models with random intercepts". Technical report, xx, http ://www.stat.umu.se/forskning/reports/glmmML.pdf.

[8] Chava, S. and Jarrow, R. A. (2004)." Bankruptcy Prediction with Industry Effects". *Review of Finance*, 8 :537–569.
[9] Durand, D.,(1941). "Risk Elements in Consumer Installment Financing, Studies in Consumer Installment Financing" Study 8, *National Bureau of Economic Research*. 88 :9 - 25.
[10] Dietsh, M., Petey, J., (2003) : " Mesure et gestion du risque de crédit dans les institutions financières", Edition La Revue Banque, Paris.
[11] Pepe, M. S. (2000). "Receiver operating characteristic methodology". *Journal of the American Statistical Association*, 95 :308–311.
[12] Press, S. J. and Wilson, S. (1978). "Choosing between logistic regression and discriminant analysis." *Journal of the American Statistical Association*,73 :699–705.
[13] Thomas, Lyn C., (2006). "Credit Scoring : The State of the Art". *FORESIGHT : International Journal of Applied Forecasting* , No. 3, pp. 33-37.
[14]Fisher, R.A., (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics*, 7 :179-188.
[15] S. Hillegeist, E. Keating, D. C. and Lundstedt, K. (2004). "Assessing the probability of bankruptcy". *Review of Accounting Studies*, 9 :5–34.20

VERS UN MODELE INTEGRATEUR DES ANTECEDENTS ET CONSEQUENCES DU RISQUE PERÇU PAR LES INVESTISSEURS SUR LE MARCHE BOURSIER TUNISIEN

Adel KARAA[a]
Maître de Conférences
Institut Supérieur de Gestion de Tunis
e-mail : adel.karaa@isg.rnu.tn
adelkaraa@yahoo.fr
Tel : 00 216 98 542 585

Azza BEJAOUI[b]
Doctorante
Institut Supérieur de Gestion de Tunis
e-mail : bjaouiazza2@yahoo.fr
bjaouiazza2@hotmail.fr
Tel : 00 216 25 935 675

[a,b]*Département des Méthodes Quantitatives & Economie, Institut Supérieur de Gestion de Tunis (I.S.G). Adresse : 41 Rue de la Liberté – Cité Bouchoucha. Le Bardo – 2000 – Tunis. Tunisie.*

Introduction

Le risque perçu joue un rôle prépondérant dans le comportement de l'être humain, surtout s'il s'agit de prendre une décision dans un contexte caractérisé par l'incertitude. C'est dans ce sens qu'on peut comprendre les propos de Ricciardi (2004) qui indique que « *si l'activité était de conduire une voiture ou investir sur le marché boursier, chaque jour, nous sommes exposés à toutes les formes du risque. Le risque peut avoir différentes significations pour différentes personnes* ». Soucieux de cette réalité, un foisonnement de recherches dans diverses disciplines a tenté d'étudier le concept du risque perçu. Tout particulièrement, en marketing, cette notion a été examiné par plusieurs études dans différents contextes tels que l'évaluation d'une marque ou d'un produit particulier (par exemple, Erdem et Swait, 2004 ; Dowling et Staelin, 1994), ou d'un service (par exemple, Bansal et Voyer, 2005 ; Murray, 1991). Le point de convergence de ces différents travaux de recherche était la mise en évidence de diverses actions entreprises par les consommateurs ayant pour but de diminuer le risque (reconnues généralement par les stratégies de réduction du risque). A titre d'exemple, la perception du risque peut déclencher chez l'individu l'activité de la recherche d'informations (Dowling et Staelin, 1994 ; Mitchell, 1992 ; Srinivasan et Ratchford, 1991). S'inscrivant dans ce courant de recherche, la présente étude tente d'examiner le rôle que pourrait jouer le risque perçu dans l'adoption des stratégies de réduction du risque perçu par une personne dans le contexte d'investissement en bourse.

A la lecture de la littérature sur le risque perçu, on peut se rendre compte que le risque perçu comprend deux composantes à savoir l'incertitude et l'importance (ou signification) des conséquences (Mallet, 2000 ; Dandouau, 2000 ; Verhage et *al*., 1990 ; Mitchell et Greatorex, 1989) engendrant, par conséquent, deux modes différents de réponses comportementales en vu de réduire le risque (Cho et Lee, 2006 ; Taylor, 1974 ; Cox, 1967a). Pour illustrer ceci, Taylor (1974) affirme que « *l'incertitude concernant le résultat peut être réduite en acquérant et traitant les informations. L'incertitude concernant les conséquences peut être traitée en réduisant les conséquences à travers la diminution du montant en jeu* ». Quoique la réduction du montant en jeu soit retenue en tant que stratégie de réduction du risque, la plupart des études en marketing se sont penchées seulement sur l'étude de la recherche d'informations. Néanmoins, Cho et Lee (2006) affirment qu'une bonne compréhension des stratégies de

réduction du risque ne peut se faire qu'à travers la prise en compte des deux éléments ensemble (c'est-à-dire la recherche d'informations et la réduction du montant en jeu).

S'intéressant à la perception du risque, la présente étude examine deux différents modes comportementaux ayant pour intention d'atténuer le risque perçu à savoir la réduction de l'incertitude à travers la recherche d'informations et la signification des conséquences *via* l'intention de réinvestissement. Outre que les stratégies de diminution du risque, la performance d'investissement est retenue comme une autre conséquence du risque perçu étant donné que c'est un facteur crucial motivant un individu à investir en bourse. En effet, un investisseur continuerait à investir davantage malgré le fait qu'il perçoit un niveau élevé du risque envers le marché boursier. Dans ce cadre, la performance d'investissement pourrait jouer un rôle de médiateur entre la perception d'investissement et l'intention de réinvestissement.

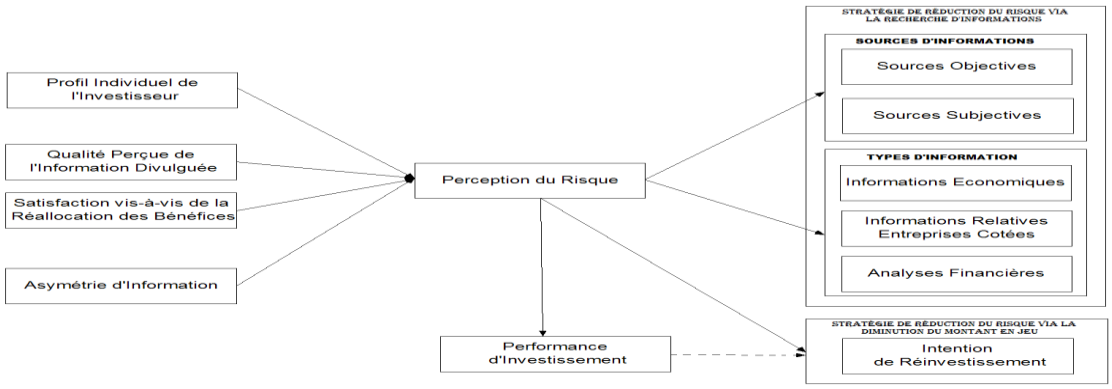
Par ailleurs, nous examinons les facteurs relatifs à l'entreprise cotée (la qualité perçue de l'information divulguée et la satisfaction de la réallocation des bénéfices), l'investisseur (le profil individuel de l'investisseur) et marché boursier (l'asymétrie d'information) comme des antécédents clés du risque perçu. Mis tous ensemble, nous proposons un modèle qui identifie les interactions entre les antécédents proposés, le risque perçu et les réponses comportementales résultantes. Ce modèle nous servira, en fait, pour tester concurremment le lien entre les facteurs influençant le risque et le risque perçu et son effet sur les deux différents modes de stratégies de réduction du risque et la performance d'investissement.

Finalement, nous pouvons noter que le modèle proposé a été analysé en se basant sur un questionnaire distribué auprès de 411 investisseurs individuels sélectionnés aléatoirement par les 24 intermédiaires présents sur le marché boursier tunisien.

1. Cadre Conceptuel

Le modèle proposé ci-dessous (figure 1) repose sur les propositions suivantes : **(1)** Le profil individuel à l'investisseur (facteur lié à l'investisseur), la qualité perçue de l'information divulguée et la satisfaction vis-à-vis de la réallocation des bénéfices (deux facteurs relatifs à la société cotée) et l'asymétrie d'information (facteur lié au marché boursier) influencent la perception du risque de l'investisseur; **(2)** La perception du risque influence le comportement ayant pour intention de gérer l'incertitude et l'importance des conséquences ; **(3)** La performance d'investissement joue un rôle de médiateur dans la relation entretenue entre la perception du risque et l'intention de réinvestissement.

Figure 1 : Modèle Conceptuel des Antécédents et Conséquences du Risque Perçu



En passant en revue les différents travaux de recherche sur le risque perçu, clair est de remarquer le manque de consensus entre les chercheurs sur la définition à donner à ce concept. Néanmoins, en revenant au travail précurseur de Bauer (1960) qui précise que « *le comportement du consommateur implique un risque dans le sens où n'importe qu'elle action du consommateur produira des conséquences qu'il ne peut pas anticiper avec certitude approximative et dont certaines sont probablement déplaisantes* », on peut comprendre que le risque perçu comporte essentiellement deux éléments : l'incertitude et les conséquences négatives (les pertes). Classiquement et suite à la contribution fondatrice de Bauer (1960), plusieurs chercheurs définissent cette notion sur la base de cette structure à deux composantes (Cho et Lee, 2006 ; Taylor, 1974 ; Cox et Rich, 1964 ; Cunningham, 1967 ; Kogan et Wallach, 1964). En guise d'exemple, Cox et Rich (1964) affirment que le risque perçu est « *une fonction de deux facteurs généraux : l'importance de l'enjeu induit par la décision d'achat et le sentiment subjectif individuel de certitude que le consommateur va perdre tout ou partie de ce qui est en jeu* ». En particulier, Cho et Lo (2006) pensent que l'évaluation du risque perçu dépend fortement de ses caractéristiques psychologiques et situationnelles vu que c'une estimation biaisée de la part d'une personne concernant une situation de décision risquée. Partageant cette opinion, Slovic (1988) affirme qu'il existe un nombre important de facteurs (ou encore antécédents) qui peuvent influencer la perception du risque d'une personne. Dans le contexte d'investissement, ils peuvent être répertorié[1] selon leur appartenance à l'investisseur, la société cotée et au marché boursier. Parmi les facteurs pouvant influencer le risque perçu par l'investisseur est son profil individuel. Ce profil est défini, en fait, en termes de deux traits de personnalité à savoir : la confiance en soi et l'optimisme. Quant aux facteurs liés aux entreprises cotées, nous avons retenu la qualité de l'information transmise au marché boursier ainsi que la satisfaction vis-à-vis de la réallocation des bénéfices par les sociétés. Finalement, l'asymétrie d'information est maintenue comme un antécédent du risque perçu qui émane du marché boursier. D'une manière plus formelle, nous avançons les hypothèses suivantes :

H1: Plus l'investisseur est confiant en soi et optimiste plus le niveau perçu du risque envers le marché boursier est faible.

[1]Cette trichotomie est originaire des études en marketing.

H2.1: Plus la qualité de l'information divulguée par les entreprises cotées est perçue comme étant bonne plus le niveau perçu du risque envers le marché boursier est faible.

H2.2: Plus l'investisseur est satisfait de la réallocation des bénéfices par des entreprises cotées plus le niveau perçu du risque envers le marché boursier est faible.

H3: Plus l'asymétrie d'information est importante sur le marché boursier plus le niveau perçu du risque est élevé.

Du coté des conséquences, le risque perçu peut être traité en utilisant deux stratégies de réduction du risque différentes. La première vise à diminuer l'incertitude *via* la recherche d'informations et la deuxième tente d'atténuer la vulnérabilité en diminuant le montant en jeu. En effet, l'investissement sur le marché boursier acquiert généralement de la part d'une personne de mettre en place une somme d'argent et supporter une quantité importante du risque. Dans ce cas, Lin et Lee (2004) précisent que la recherche d'informations s'avère une activité essentielle que chaque investisseur doit l'accomplir en vu de prendre ses décisions d'investissement. Position partagée par Loibl et Hira (2009) qui indiquent que le risque d'une perte financière importante et les coûts élevés engendrés par la révision d'une mauvaise décision d'investissement motivent les investisseurs à rechercher l'information. Chercher l'information revient à consulter les sources (objectives et subjectives) et types d'information (informations économiques, les informations relatives aux entreprises cotées, les analyses financières). De l'autre côté, la perception du risque pourrait être réduite en restreignant le montant en jeu ou encore la vulnérabilité associée aux résultats défavorables. Un tel comportement vise à diminuer l'impact des conséquences négatives qui résulte d'une prise de décision. Dans ce cadre, le comportement clé serait de réviser l'intention de réinvestissement sur le marché boursier. Mise à part ces stratégies, une personne investissant sur le marché boursier cherchera inévitablement à réaliser des gains. A ce propos, Swanson et Lin (2005) ajoutent que la performance d'investissement est le résultat du comportement de l'investisseur sur le marché boursier. Par ailleurs, en étudiant les mécanismes psychologiques des comportements d'investissement risqué auprès de 1547 investisseurs individuels présents sur les marchés boursiers chinois, Wang et *al*. (2006) concluent qu'il existe une relation positive entre la perception du risque et la performance d'investissement. La rétention de ce facteur dans la présente étude vise à examiner si la performance d'investissement joue un rôle de médiateur dans l'effet de la perception du risque sur l'intention de réinvestissement. Ainsi, nous formulons les hypothèses suivantes :

H4: Plus la perception du risque de l'investisseur est élevée plus il consultera diverses sources d'informations.

H5 : Plus la perception du risque de l'investisseur est élevée, plus différents types d'information vont être utilisés.

H6 : Une perception du risque élevée de la part de l'investisseur mènera une faible intention de réinvestissement sur le marché boursier.

H7 : La performance d'investissement joue un rôle de médiateur entre la perception du risque l'intention de réinvestissement.

2. Présentation de la Méthodologie de Recherche

L'ensemble des données utilisé afin d'examiner le cadre conceptuel proposé a été collecté à l'aide d'un questionnaire auto-administré auprès des investisseurs individuels présents sur la Bourse des Valeurs Mobilières de Tunis (BVMT). Les individus sondés sont, en fait, choisis aléatoirement par les vingt quatre intermédiaires en bourse.

Environ 600 questionnaires ont été distribués parmi lesquels 487 ont été récupérés, enregistrant un taux de réponse de 81%. Néanmoins, lors de dépouillement, une soixante dizaine de questionnaires a été ôtée à cause de l'existence d'un nombre élevé de non réponses. Finalement, 411 questionnaires sont jugés, de notre part, exploitables pour être analysées.

❖ *Typologie des Répondants*

80.3% des investisseurs individuels étaient des hommes alors que les femmes ne représentaient que 19.7%. La tranche d'âge la plus importante dans l'échantillon est celle des 30-39 ans (37%), suivie par celle des 40-49 ans (27,5%) et ensuite celle des 20-29 ans (20,2%). Finalement, la tranche d'âge la moins importante s'avère celle des plus que 60 ans (4,1%). Par ailleurs, La majorité des répondants déclarent être mariés avec enfants, ceci à un niveau de 54.4% de l'échantillon étudié. En outre, l'échantillon se compose des cadres supérieurs, des cadres moyens, des employés, des indépendants, des retraités, des étudiants et des chômeurs à hauteur respectivement de 40.1%, 28.6%, 12.1%, 12.1%, 4.7%, 2.2% et 0.2%. Finalement, les résultats relatifs aux tranches de revenu par investisseur dévoilent que 26.5% des répondants touchent entre 1500 et 2000 DT alors que 5.3% seulement ont un revenu inférieur à 500 DT.

❖ *Mesures*

Plusieurs mesures sont construites afin de mesurer les différents concepts du modèle proposé dans cette étude. En particulier, le risque perçu a été mesuré en se référant au paradigme psychométrique développé par plusieurs chercheurs en psychologie (par exemple, Slovic, 1987). En effet, les résultats trouvés à partir des travaux de recherche psychométriques montrent que les individus évaluent n'importe quel événement risqué en se basant principalement sur deux dimensions orthogonales à savoir : la familiarité (*familiarity)* et la maîtrise (*controllability*). Ainsi, nous utilisons ces deux facteurs en vu de quantifier le risque perçu par l'investisseur. En termes opérationnels, la familiarité et la maîtrise ont été mesurées par un ensemble d'*items* représentant les instruments nécessaires à l'investissement (par exemple, les ratios financiers, la fluctuation du taux d'intérêt…) mais tel que les deux dimensions du risque perçu pourraient se différencier l'une de l'autre. Une fois mesurées, nous pensons que le degré de familiarité et maîtrise des outils servant à l'investissement en bourse déterminent le risque perçu par l'investisseur. En d'autres termes, ces deux dimensions ne représentent pas des manifestations du risque perçu mais à *contrario* sont constructives (ou encore formatives) de ce construit. Ainsi, nous présumons que le modèle de mesure du risque perçu est un modèle formatif.

3. Analyse des Données et Résultats

L'analyse des données s'est déroulée en trois étapes. D'abord, des Analyses en Composantes Principales (ACP) réalisées sur l'ensemble des données collectées (N=411) *via*

la version 17 du logiciel SPSS ont servi à purifier les échelles de mesures retenues dans cette étude. Ensuite, une Analyse Factorielle Confirmatoire (AFC) avec la version 18 du logiciel AMOS a été employée en vu de vérifier la structure factorielle déjà trouvée à partir de la phase exploratoire. Finalement, pour examiner les relations hypothétiques proposées dans le modèle, nous avons eu recours à la modélisation par les équations structurelles.

Les résultats montrent, tout d'abord, que les deux variables latentes, Familiarité et Maîtrise, sont formatives du construit 'Risque Perçu'. Vu que le modèle de mesure du risque perçu est un modèle formatif, les facteurs influençant la perception du risque vont être, par la suite, considérés comme des antécédents de la familiarité et la maîtrise. Par ailleurs, les résultats indiquent que le profil individuel de l'investisseur est lié positivement à la familiarité et la maîtrise des outils d'investissement. Ces résultats confirment le fait qu'un investisseur confiant en soi et optimiste croît connaître et maîtrise les instruments nécessaires à l'investissement. Il percevra, par conséquent, un niveau de risque faible envers le marché boursier. Nous constatons également que la qualité perçue de l'information livrée par les entreprises cotées et la satisfaction vis-à-vis de la réallocation des bénéfices entretiennent, chacune des deux, une relation positive et significative avec la familiarité et la maîtrise. Ainsi, on peut dire qu'elles renforcent, toutes les deux, chez l'investisseur le sentiment qu'il est familier et maîtrise les instruments d'investissement. Dés lors, sa perception du risque envers le marché sera d'autant moins élevée. Contrairement à ce qui a été prévu, les résultats montrent que plus les participants du marché ne partagent pas la même information plus le degré de familiarité et maîtrise de l'investisseur est élevée. Ce résultat pourrait être expliqué par le fait que l'investisseur estimant être plus informé que son autrui présent sur le marché tend à se sentir plus familier et maîtrise les outils d'investissement. Par conséquent, il percevra un niveau du risque plus faible envers le marché boursier.

Du côté des conséquences, nous remarquons que plus la perception du risque de l'investisseur est élevée plus il consultera diverses sources d'informations. Ainsi, le risque perçu envers le marché mènera l'investisseur à ne pas préférer une source d'informations particulière (source objective ou subjective). A *contrario*, il cherchera différentes sources d'informations en vu de réduire davantage sa perception du risque. En outre, les résultats dégagés suggèrent qu'en faisant face à un niveau élevé du risque perçu l'investisseur aura tendance à utiliser différents types d'information (informations économiques, analyses financières, informations relatives aux sociétés cotées). En d'autres termes, le participant au marché boursier utilise divers types d'information afin d'atténuer son risque perçu. Nous concluons ainsi que le lien entre la perception du risque et l'activité de la recherche d'informations peut être établi dans le contexte d'investissement. Néanmoins, nous remarquons que la perception du risque a un effet significatif et positif, au lieu d'être négatif, sur l'intention de réinvestissement ce qui amène à rejeter l'hypothèse H6. Pour finir, les résultats montrent que la performance d'investissement peut jouer un rôle d'un médiateur partiel entre la perception du risque et l'intention de réinvestissement.

Conclusion

Dans la présente étude, le modèle proposé met en évidence les antécédents et les conséquences du risque perçu dans le contexte d'investissement en bourse. Les résultats

prouvent que la maîtrise joue un rôle crucial dans la formation de la perception du risque de l'investisseur. Par ailleurs, nous constatons que le profil individuel de l'investisseur, les facteurs liés à l'entreprise cotée de même que celui lié au marché boursier ont un impact sur le risque perçu. Une étude future pourrait examiner l'effet d'autres variables sur la perception du risque telles que l'expérience d'investissement et les caractéristiques sociodémographiques (âge, genre, revenu…). De même, nous concluons que les investisseurs cherchent à atténuer le risque perçu envers le marché boursier à travers la recherche d'informations. L'attraction des rendements dégagés renforce chez les investisseurs l'intention de réinvestir bien qu'ils perçoivent un risque envers le marché boursier. Ainsi, la spécificité du contexte (c'est-à-dire l'investissement en bourse) mène l'investisseur à ne pas adopter la deuxième stratégie de réduction du risque à savoir la diminution du montant en jeu.

Bibliographie

[1] Bansal, H.S. et Voyer, P.A. (2000) World-of-Mouth Processes Within Services Purchase Decision Context, *J Serv Res*, 3(2), 166-78.

[2] Bauer, R.A. (1960) Consumer Behavior as Risk Taking. in R.S. Hancock, ed., *Dynamic marketing for a changing world* (pp. 389-398). Chicago: American Marketing Association.

[3] Cho, J. et Lee, J., (2006) An Integrated Model of Risk and Risk-Reducing Strategies, *Journal of Business Research*, vol. 59, issue 1, 112-120.

[4] Cox, D.F (1967a) Risk Taking and Information Handling in Consumer Behavior, *Harvard University Press*, Cambridge, MA.

[5] Cox, D.F. et Rich, S.U. (1964) Perceived Risk and Consumer decision Making-The Case of Telephone Shopping, *Journal of Marketing Research*, 1, 32-39.

[6] Cunningham, S.M. (1967) The Major Dimensions of Perceived Risk, Risk Taking and Information Handling in Consumer Behavior. D.F. Cox E., Boston, *Harvard University Press*, 82-108.

[7] Dandouau, J.C. (2000) Le Comportement de Recherche d'Informations des Différents Profils de Risque Perçu selon la Nature de l'Achat, *in les Actes des liers ateliers de recherche de l'AFM*, « Percevoir, identifier et gérer le risque en marketing », La Sorbonne, Paris, 133-151.

[8] Dowling, G.R. et Staelin, R. (1994) A Model of Perceived Risk and Intended Risk Handling Activity, *J Consum Res*, 21, 119– 134.

[9] Erdem, T. et Swait, J. (2004), Brand Credibility, Brand Consideration and Choice, *J Consum Res*, 31(1).

[10] Kogan, N., Wallach, M.A. (1964), Risk Taking: A Study in Cognition and Personality. Holt, New-York, Rinehart & Winston.

[11] Lin, Q. et Lee, J. (2004) Consumer Information Search When Making Investment Decisions, *Financial Services Review*, 13, 319-332.

[12] Loibl, C. et Hira, T.K. (2009) Investor Information Search, *Journal of Economic Psychology*, 30, 24-41.

[13] Mallet, S. (2000) Le Concept de Risque Perçu: Composantes, Antécédents et Proposition de Recherche, *in les Actes des liers ateliers de recherche de l'AFM*, « Percevoir, identifier et gérer le risque en marketing », La Sorbonne, Paris, 46-62.

[14] Mitchell, V.-W. (1992) Understanding Consumer's Behavior: Can Perceived Risk Theory Help?, *Management Decision*, 30, 3, 26-31.
[15] Mitchell, V.-W. et Greatorex, M., (1989) Risk Reducing Strategies Used in the Purchase of Wine in the UK, *European Journal of Marketing*, 23 (9), 31-46.
[16] Murray K.B. (1991) A Test of Services Marketing Theory: Consumer Information Acquisition Activities. *J Mark*, 55, 10–25.
[17] Ricciardi, V. (2004) A Behavioral Finance Study: An Investigation of the Perceived Risk for Common Stocks by Investment Professionals (financial analysts vs. financial planners), *Dissertation*. (Doctor of Business Administration in Finance): Golden Gate University.
[18] Slovic, P. (1987) Perception of Risk, *Science*, 236, 280-285.
[19] Slovic, P. (1988) Risk perception. In C. C. Travis (Ed.), *Contemporary issues in risk analysis: Vol. 3: Carcinogen risk assessment*, 171-181. New York: Plenum.
[20] Srinivasan, N. et Ratchford, B.T. (1991) An Empirical Test of a Model of External Search for Automobile, *J Consum Res*, 18(2), 233-243.
[21] Swanson, P.E et Lin, A.Y. (2005) Trading behavior and Investment Performance of U.S. Investors in Global Equity Markets, *Journal of Multinational Financial Management*, 15, 99-115.
[22] Taylor, J. (1974) The Role of Risk in Consumer Behavior, *Journal of Marketing*, 38, 2, 54-60.
[23] Verhage, B.J., Yavas, U. et Green, R.T. (1990) Perceived Risk: A Cross-Cultural Phenomenon, *International Journal of Research in Marketing*, 7, 297-303.
[24] Wang, X.L, Shi, K. et Fan, H.X. (2006) Psychological Mechanisms of Investors in Chinese Stock Markets, *Journal of Economic Psychology*, 27, 762-780.

UNE MÉTHODE DE TRAITEMENT DES REFUSÉS DANS LE PROCESSUS D'OCTROI DE CRÉDIT

Asma Guizani[1] & Salwa Ben Ammou[2] & Gilbert Saporta[3]

[1]Institut Supérieur de Gestion de Sousse, rue Abdlaaziz il Behi . Bp 763. 4000 Sousse Tunisie.
[2]Faculté de Droit & des Sciences Economiques et Politiques de Sousse Cité Erriadh - 4023 Sousse Tunisie.
[3]Laboratoire Cédric - CNAM, 292 rue Saint Martin, 75141 Paris cedex 03, France.

Abstract

The object of our paper is to build a credit scoring model based on a representative sample of the total population (accepted + rejected) to remedy the problem of selection bias, PLS-DA regression (PLS Discriminant Analysis) and canonical discriminant analysis are used for their simplicity and efficacy.

Résumé

On présente deux modèles de scoring construits sur la base d'un échantillon représentatif de la population globale (acceptés + refusés) pour remédier au problème du biais de sélection. Les modèles adoptés dans notre cas c'est la régression PLS-DA et l'analyse factorielle discriminante.

Thèmes : Apprentissage et classification, Modèles pour les assurances et les finances.
Mots clés : crédit scoring, régression PLS-DA, analyse factorielle discriminante, augmentation simple, courbe ROC.

1. Introduction

Le risque de défaillance de l'emprunteur a toujours constitué, pour une banque, le cœur même de la problématique de l'analyse financière qui accompagne toute demande de crédit. Disposer de modèles statistiques pour prédire la défaillance est donc devenue primordial pour une banque, surtout dans le contexte actuel de renforcement du contrôle des risques bancaires et le respect de la nouvelle réglementation prudentielle préconisé par les accords « Bâle II » et « Bâle III ».
Le crédit scoring est un outil fondamental de prévision des risques basé sur les caractéristiques du demandeur de prêt. À partir d'un échantillon de dossiers acceptés dont la qualité est connue, on calcule une note, le score, dont on déduit la probabilité de défaut. Cette probabilité ne peut être estimée que pour les dossiers acceptés. On ne peut donc l'estimer pour les demandeurs rejetés dés le départ, (données incomplètes) ce qui peut conduire à des estimations incorrectes (biais de sélection [2]) et à un éventuel manque à gagner pour la

banque si le client rejeté a priori alors qu'en réalité c'est un bon payeur qui n'aurait pas fait défaut si un crédit lui avait été octroyé.
C'est la problématique de la classification semi-supervisée où l'on dispose à la fois d'un ensemble de données étiquetées (dossiers acceptés) et d'un ensemble de données non étiquetées (dossiers refusés). Le grand intérêt de la classification semi-supervisée est de pouvoir combiner l'information contenue dans des données étiquetées et celle contenue dans les données non étiquetées afin d'atteindre des taux de classification plus élevé.
Le traitement des refusés (reject inference en anglais) tente de remédier à ce problème et de corriger biais de sélection en réintégrant les dossiers refusés à l'échantillon initial et par la suite rendre ce dernier représentatif de la population globale (admis+refusés).
Dans la section 2, nous présentons les techniques classiques de traitement des refusés.
La méthode d'inférence de rejet, adoptée dans notre cas, est présentée dans la section 3. Ensuite, nous mettons en application, dans la section 4, cette méthode pour répondre à notre problématique et nous comparons la performance des modèles de score obtenus. Enfin, la section 5 est consacrée aux conclusions et perspectives de recherches pour la mise en ouvre d'autres méthodes et la définition du modèle le plus performant.

2. Panoramas des techniques de traitement des refusés

Parmi les nombreuses méthodes de traitement des refusés dans la littérature, mentionnons :

2.1 L'augmentation

Cette méthode consiste à construire d'abord un score d'acceptation qui prévoit la probabilité d'être accepté parmi la population globale. On applique ce modèle à la population toute entière et on sépare cette dernière en intervalles ou bandes de score selon le critère de notre choix. Á chaque intervalle de score, on définit un poids, chaque dossier accepté est pondéré par ce poids et un modèle de score de défaut est construit sur les acceptés ainsi pondérés.

Intervalle de score	**Nombre d'acceptés**	**Nombre de refusés**	**Poids**
1	A1	R1	(A1+R1)/A1
…	**…**	**…**	**…**
N	An	Rn	(An+Rn)/An

Tableau n°1 : Table de calcul de poids [7]

2.2 Parceling [7]

On construit un modèle de score sur les acceptés, ensuite on départage la population en intervalles de score puis on calcule, sur chaque intervalle, le nombre de défaut et de non défaut des dossiers acceptés et le nombre total des refusés. On applique une hypothèse de taux de défaut sur les refusés (qui nous donne le nombre de défaut et non défaut dans la population des refusés). On définit ensuite, au hasard, les refusés de chaque intervalle en deux classes défaut/non défaut tout en respectant le nombre de défaut/non défaut calculé dans chaque intervalle. On aboutit ainsi à la constitution du « augmented data set » (on regroupe les acceptés initiaux et les refusés calculés à partir de l'hypothèse de taux de défaut) sur le quel on va construire notre modèle de score.

2.3 L'extrapolation [7]

On construit un modèle de score sur les acceptés et on l'applique à tous les dossiers. L'inconvénient de cette méthode, c'est qu'il y a un biais qui peut être positif ou négatif.

2.4 Reclassification « augmented data set » [7]

On construit un modèle de score de défaut sur les acceptés et on l'applique aux dossiers rejetés. Ces derniers sont classifiés en deux catégories défaut et non défaut.
On construit alors le « augmented data set » qui consiste à ajouter aux dossiers acceptés initiaux (étiquetés défaut/non défaut) les dossiers rejetés étiquetés inférés (défaut/non défaut).
Finalement, on produit le modèle de score sur le « augmented data set ». Cette méthode aboutit donc à l'hypothèse implicite que la distribution des défaut/non défaut est la même dans les populations d'acceptés et de rejetés, ce qui est en réalité faux.

2.5 Groupe de contrôle [7]

Cette méthode consiste à accepter tous les dossiers d'un groupe de contrôle représentatif de la population complète. On construit par la suite notre modèle de score sur cet échantillon départagé en deux catégories défaut et non défaut.

3. La méthode de l'augmentation simple [5]

Afin de résoudre le problème du biais de sélection, nous utilisons la méthode de l'augmentation simple qui se résume en les étapes suivantes :

- Etape 1 : construire un modèle de score sur la base d'un échantillon composé des dossiers acceptés seulement qui sont étiquetés en bon et mauvais payeur (c'est notre échantillon d'apprentissage).
- Etape 2 : appliquer le modèle établi sur les refusés et déterminer le taux de défaut de ces derniers (c'est le principe même du crédit scoring adopté par les banques mais qui est ici appliquer sur des dossiers qui devront normalement être refusés).
- Etape 3 : étiqueter les dossiers refusés par bon ou mauvais selon le taux de défaut.
- Etape 4 : une fois l'échantillon des refusés est défini en bon et mauvais payeur, il sera par la suite réintégrer à l'échantillon d'apprentissage de l'étape 1 pour reconstruire un nouveau modèle non biaisé sur la base cette fois ci d'un échantillon représentatif de la population globale (acceptés et refusés). Ce qui représente un risque considérable à prendre par la banque et qui peut lui coûter cher.

Dans notre cas, nous adoptons deux méthodes pour la construction du modèle du score, la régression PLS-DA (Partial Least Squares Discriminant Analysis) [6] et l'analyse factorielle discriminante (AFD) [3]. La variable à prédire étant binaire (1 ou 0) selon la qualité du dossier. On sait en effet que l'analyse discriminante linéaire est dans ce cas identique à une régression multiple. Le score est alors défini comme la combinaison linéaire des variables obtenues par la régression et on étudie les performances de prédiction en faisant varier le terme constant (ou seuil).

4. Modélisation

Les données utilisées proviennent de la compétition PAKDD 2010 (Pacific-Asia Conference on Knowledge Discovery and Data mining qui s'est déroulée en Inde du 21 au 24 juin 2010)[1]. Ce sont des dossiers d'octroi de crédit d'une banque brésilienne sur une période s'étalant de 2006 à 2009. Notre objectif est de construire un modèle qui sert à étiqueter les clients en bon et mauvais payeur. La variable dépendante est donc une variable binaire qui indique si le client a fait défaut pour une période de 60 jours pendant la première année du crédit (Y = 1), sinon Y = 0. Nous comptons 22 variables explicatives dont 7 sont des variables quantitatives et les 15 autres sont des variables qualitatives. Dans ce cadre de mélange de données qualitatives et quantitatives, on a transformé chaque variable qualitative à r modalités en r variables numériques indicatrices de chaque modalité. Nous disposons en tout de 15000 dossiers dont 10000 dossiers étiquetés en bon et mauvais payeur et ne comportant que les dossiers acceptés (ils seront utilisés pour la construction de notre modèle) et 5000 dossiers non étiquetés comportant les dossiers acceptés et refusés et seront utilisés pour la prévision.
Afin de remédier au problème de traitement des refusés, nous avons appliqué la méthode de l'augmentation simple (évoquée précédemment) qui consiste tout d'abord à construire un modèle de score sur la base des 10000 dossiers (acceptés seulement qui sont étiquetés en bon et mauvais payeurs : c'est notre échantillon d'apprentissage) en utilisant une régression PLS-DA que nous comparons par la suite au modèle obtenu par l'analyse factorielle discriminante. Une fois le modèle établi (quelque soit par une régression PLS-DA ou une analyse factorielle discriminante), il sert à prédire et étiqueter les 5000 dossiers supplémentaires (acceptés+refusés : non étiquetés) en bon et mauvais payeur, la règle de décision bancaire en matière d'étiquetage des dossiers s'appuie sur le score obtenu à partir du modèle. Donc, à un score Z calculé inférieur à Z_c (score limite ou seuil) serait associée l'étiquette bon payeur (Y=0), à un score Z calculé supérieur à Z_c serait associée l'étiquette mauvais payeur (Y=1). Une fois, les dossiers étiquetés (selon le modèle construit initialement), on les réintègre aux 10000 premiers dossiers pour reconstruire un nouveau modèle de score non biaisé car il prend en considération un échantillon représentatif de la population globale.
Pour étudier la performance de nos modèles, nous avons utilisé la courbe ROC (Receiver Operating Characteristics) qui relie la proportion de vrais positifs (bons dossiers classés tels) à la proportion de faux négatifs (mauvais dossiers classés bons) lorsqu'on fait varier le seuil du score d'acceptation. L'aire sous la courbe (ou Area Under the Curve – AUC) est un indice synthétique calculé pour la courbe ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. L'AUC appartient à l'intervalle [0,1], un modèle est considéré comme idéal si l'AUC est égal à 1. La figure 1 représente les courbes ROC pour les deux modèles (régression PLS-DA et analyse factorielle discriminante) avant réintégration des 5000 dossiers supplémentaires.
Nous remarquons, d'après cette figure, que le modèle construit avec l'analyse factorielle discriminante (AUC=0,603) est plus performant que celui construit avec la régression PLS-DA (AUC=0,595), ce qui n'est cependant pas une très bonne performance.

[1] http://sede.neurotech.com.br/PAKDD2010/

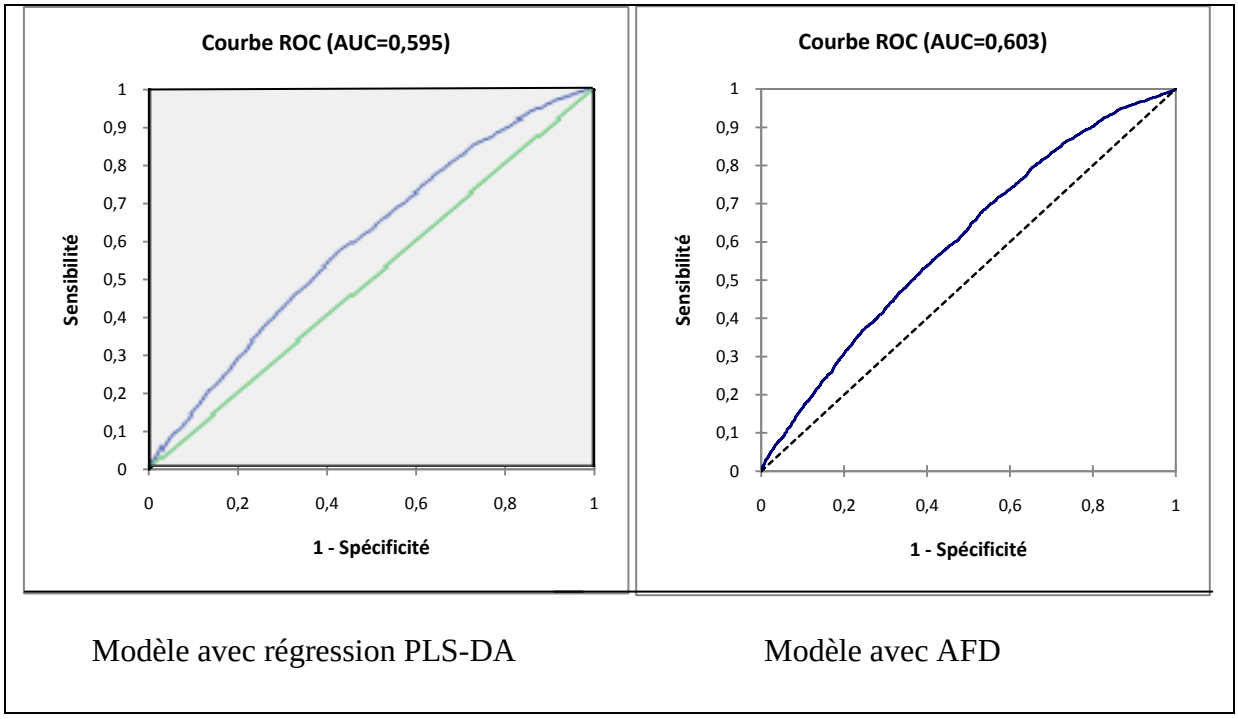


Figure 1 : Courbes ROC avant réintégration des dossiers supplémentaires

La figure 2 représente les courbes ROC des deux modèles une fois les 5000 dossiers réintégrés.

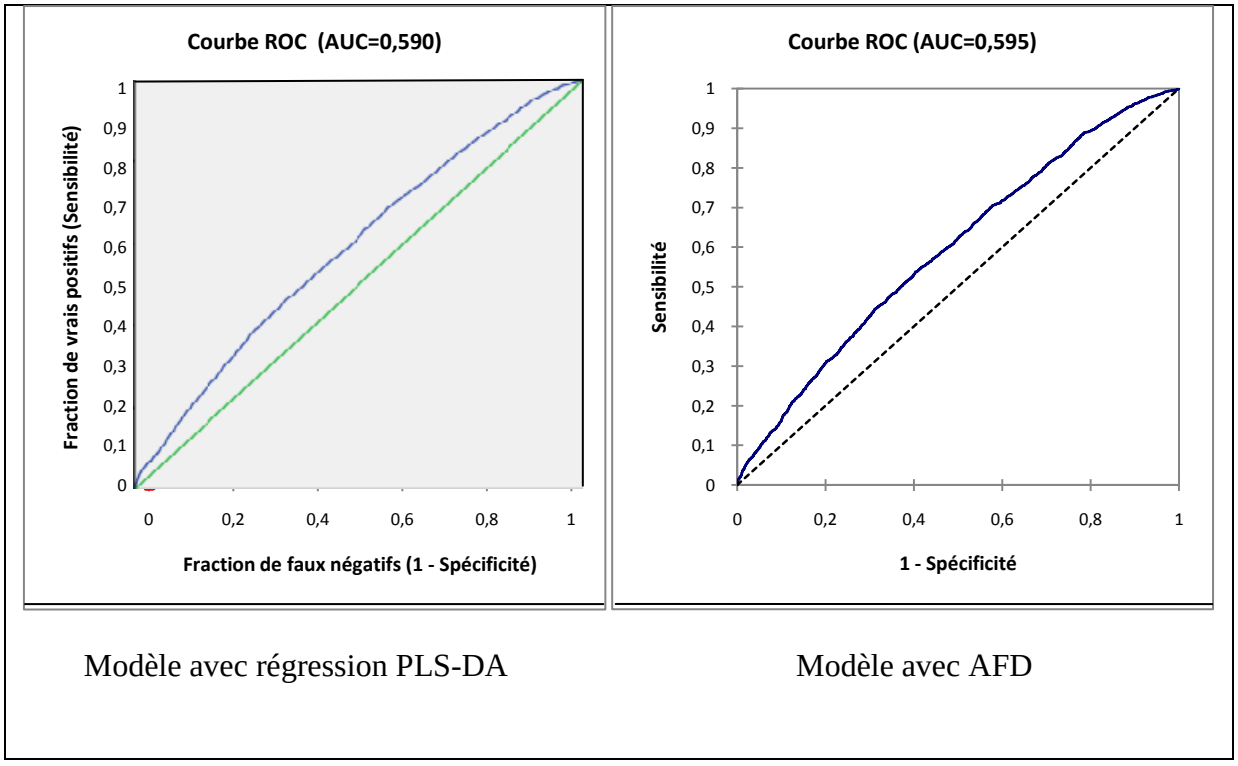


Figure 2 : courbes ROC après réintégration des 5000 dossiers supplémentaires

Nous constatons que la valeur de l'AUC, pour les deux modèles construits avec la régression PLS-DA et l'analyse factorielle discriminante, reste presque la même pour les deux cas de figure (avant et après réintégration des dossiers supplémentaires). Elle passe de 0.595 à 0.590

pour le premier modèle et de 0.603 à 0.595 pour le deuxième modèle, mais il reste toujours plus performant que celui de la régression PLS-DA.
Pour valider nos deux modèles, nous avons calculé, pour l'analyse factorielle discriminante, le taux de bon classement des individus. Ce qui aboutit aux résultats suivants : 55.2% des observations d'origine sont correctement classées et 54.9% des individus ayant subi une validation croisée sont bien classées, ce qui révèle que notre modèle discriminant est moyennement significatif. Quant' à la régression PLS-DA, nous avons appliqué le modèle sur l'échantillon de prévision pour déterminer la classe de chaque individu, la valeur de l'AUC est de 0.588.

5. Conclusions et perspectives

Le traitement des refusés dans le processus d'octroi de crédit est une méthode qui permet de remédier au problème du biais de sélection en réintégrant les dossiers refusés par la banque dans l'échantillon qui sert à construire le modèle de score.
Nous nous sommes basés sur la méthode de l'augmentation simple comme méthode de l'inférence de rejet pour aboutir à un modèle de score construit sur la base d'un échantillon représentatif de la population globale. Nous avons mis en œuvre deux techniques statistiques pour construire notre modèle de score, à savoir la régression PLS-DA et l'analyse factorielle discriminante. Cette dernière s'avère plus performante que la première. L'objectif pour la suite est de mettre en œuvre d'autres méthodes de réintégration des refusés et de comparer les performances de chacune pour déterminer la méthode la plus efficace à corriger le biais de sélection et ainsi d'éviter à la banque le manque à gagner en refusant un dossier qui peut s'avérer par la suite un bon payeur.

Bibliographie

[1] Anderson, R. (2007) *The credit scoring toolkit theory and practice for retail credit risk management and decision automation*. Oxford University Press, New York.

[2] Banasik, J. et Crook, J. (2007*) Reject inference, augmentation, and sample selection*. European Journal of Operational Research, 183, 1582-1594.

[3] Bardos, M. (2001) *Analyse discriminante ; application au risque et scoring financier*. Dunod, Paris.

[4] Bardos, M. (2005) *Les scores de la Banque de France : leur développement, leurs applications, leur maintenance*. Bulletin de la Banque de France, 144, 63-73.

[5] Siddiq, N. (2006) *Credit risk scorecards developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc., New Jersey.

[6] Tenenhaus, M. (1998) *La régression PLS*. Editions Technip, Paris

[7] Viennet, E., et Fogelman Soulié, F. (2007) *Le traitement des refusés dans le risque crédit*. Revue des Nouvelles Technologies de l'Information (RNTI-A-1), 23-45.

Index