

# Séminaire de statistique ENSAI

Année 2005-2006

**Responsable Valentin Patilée**

## Vendredi 23 juin 2006

11h00 à 11h50 : **Olivier LOPEZ** (ENSAI)

*Tests non paramétriques d'adéquation en présence de censure.*

**Résumé.** Le modèle de régression linéaire en présence de censure sur la variable expliquée a suscité beaucoup d'intérêt au sein de l'analyse de survie. Pour prendre en compte la censure, deux des approches les plus courantes consistent soit à corriger les observations de la variable expliquée, soit à procéder à l'estimation par moindres carrés pondérés. A notre connaissance, le problème de test d'adéquation du modèle linéaire en présence de censure, ou plus généralement d'un modèle paramétrique, contre une alternative non-paramétrique a été très peu exploré. Dans ce travail, nous proposons deux tests non-paramétriques impliquant des estimateurs à noyau et dont la validité repose sur l'utilisation d'intégrales de l'estimateur de Kaplan-Meier. Nos deux procédures de validation de modèle transposent les approches d'estimation mentionnées au problème de tests.

## Vendredi 26 mai 2006

11h00 à 11h50 : **Winfried STUTE** (Mathematisches Institut, Justus-Liebig-Universität Gießen)

*Model diagnosis for parametric regression in high-dimensional spaces.*

**Résumé.** ([abstract1](#))

12h00 à 12h50 : **Winfried STUTE** (Mathematisches Institut, Justus-Liebig-Universität Gießen)

*Nonparametric comparison of two regression functions.*

**Résumé.** ([abstract2](#))

## Vendredi 21 avril 2006

11h00 à 11h50 : **Thierry KAMIONKA** (CREST Paris, Lab. de Microéconométrie)

*The Returns to Seniority in France (and Why are They Lower than in the United States?)*

**Résumé.** Dans cet article, nous estimons sur données individuelles françaises un modèle dynamique à trois équations: salaire, participation et mobilité. Ce modèle permet de distinguer entre hétérogénéité individuelle inobservable et dépendance d'état. Nous estimons ce modèle en utilisant une approche bayésienne à partir de données de panel observées sur longue période (1976-1995). Nos résultats montrent clairement que les rendements de l'ancienneté sont faibles et, pour certains niveaux d'éducation, il sont proches de zéro. La spécification retenue est proche de celle employée par Buchinsky, Fougère, Kramarz et Tchernis (2002). Ces auteurs trouvent qu'aux Etats-Unis, les rendements de l'ancienneté sont relativement importants. Les résultats restent valides lorsque l'on utilise la méthode proposée par Altonji et Williams (1992). Il s'avère que les différences de rendements obtenues découlent de la mobilité inter-entreprises. Le modèle de Burdett et Coles (2003) nous permet de comprendre ce résultat. Dans un pays comme la France où la mobilité est faible, les entreprises ont moins intérêt à compenser les salariés ayant des anciennetés importantes parce qu'ils ont tendance à passer une grande partie de leur carrière dans la même entreprise. Cela est vrai même lorsque les individus possèdent une quantité substantielle de capital humain spécifique. Au contraire, dans un pays comme les Etats-Unis où la mobilité est importante, des rendements élevés de l'ancienneté ont un effet incitatif important et les entreprises sont conduites à payer la prime associée au capital humain spécifique afin d'éviter de perdre leurs salariés les plus productifs. (Travail en collaboration avec Magali Beffy, Moshe Buchinsky, Denis Fougère et Francis Kramarz.)

## Vendredi 31 mars 2006

11h00 à 11h50 : **Pierre DEL MORAL** (Univ. Nice Sophia-Antipolis).

*Interprétations Probabilistes des Formules de Feynman-Kac*

**Résumé.** Les formules de Feynman-Kac offrent une description probabiliste précise d'un certain nombre de problèmes non linéaires issus de diverses disciplines scientifiques connexes. Nous aborderons dans cet exposé, quelques domaines d'applications relevant de l'analyse d'événements rares, de la physique des particules, de la biologie macro-moléculaire, en enfin du traitement du signal, et du filtrage non linéaire.

Nous soulignerons les problèmes mathématiques sous jacents associés à l'analyse asymptotique, et à l'estimation numérique, de lois

conditionnelles complexes dans des espaces de grandes dimensions, de probabilités de temps de sortie, d'absorption, et d'évènements rares, ainsi qu'à l'analyse spectrale de semigroupes de Feynman-Kac-Schrödinger. La seconde partie de cet exposé sera consacré à la résolution numérique, et plus précisément aux interprétations de ces formules en terme de systèmes de particules en interaction, et de modèles d'arbres généalogiques.

12h00 à 12h50 : **Fabien CAMPILLO** (IRISA, Rennes)  
*Analyse bayésienne de modèles d'évolution de ressources naturelles*

**Résumé.** Les méthodes de Monte Carlo par Chaîne de Markov (MCMC) permettent d'approcher les lois a posteriori de problèmes statistiques avec observations partielles et bruitées. Dans le cas de modèles complexes, par exemple dynamique, ces méthodes peuvent toutefois s'avérer très lentes à converger. Une alternative propose de combiner l'aspect adaptatif de MCMC et la capacité de l'échantillonnage d'importance à proposer des échantillons i.i.d. Ces méthodes s'apparentent à des algorithmes de Monte Carlo séquentiels. Nous intéressons ici à des modèles paramétriques d'évolution de ressources naturelles. Ils sont le plus souvent non linéaires et doivent être calés sur un très petit nombre d'observations. (travail joint avec Vivien Rossi)

## Vendredi 10 mars 2006

11h00 à 11h50 : **Ingrid VAN KEILEGOM** (Univ. catholique de Louvain, Belgium).  
*Estimation of a Semiparametric Transformation Model*

**Résumé.** We propose consistent estimators for transformation parameters in semiparametric models. The problem is to find the optimal transformation into the space of models with a predetermined structure. Special cases are searching for the transformation to yield an additive or a multiplicative nonparametric model. We do not focus on a particular non-parametric estimator, but give results for the estimation of the transformation when the rest of the model is estimated non- or semi-parametrically and fulfills some consistency conditions. For the estimation of the transformation parameter are proposed two methods: minimizing the mean squared distance from independence or maximizing a profiled likelihood function. First is discussed the problem of identification of such models. We then state results for a general class of nonparametric estimators. Finally we give some particular examples for nonparametric estimation of transformed separable models. The theoretical results as well as the small sample performance are studied by several simulation exercises. (This work is joint work with Oliver Linton and Stefan Sperlich.)

12h00 à 12h50 : **Daniela COCCHI** (Università di Bologna, Italy).  
*Recovering information from synthetic air quality indices.*

**Résumé.** Synthetic indices are often used to condense complex situations into a single figure. However, this condensing process risks losing potentially useful information, especially when the index is to be utilized by public decision-making bodies. The present study proposes a general strategy, combining a number of different methods, designed to recover information from air-quality indices: graphical methods to reconstruct the composition of pollution, multinomial logit analysis to study the influence of meteorological covariates on air-quality indices, and finally, a probability distribution for the index itself as a basic tool with which to interpret the index's crucial values.

## Vendredi 10 février 2006

11h00 à 11h50 : **Fabienne COMTE** (Univ. Paris 5).  
*Estimation non paramétrique par moindres carrés pénalisés des coefficients d'une diffusion. (PDF)*

**Résumé.** We consider a one-dimensional diffusion process  $(X_t)$  which is observed at  $n+1$  discrete times with regular sampling interval  $\Delta$ . Assuming that  $(X_t)$  is strictly stationary, we propose nonparametric estimators of the drift and diffusion coefficients obtained by a penalized least square approach. Our estimators belong to a finite dimensional function space whose dimension is selected by a data-driven method. We provide non asymptotic risk bounds for the estimators. When the sampling interval tends to zero while the number of observations and the length of the observation time interval tend to infinity, we show that our estimators reach the minimax optimal rates of convergence. Numerical results based on exact simulations of diffusion processes are given for several examples of models and enlight the qualities of our estimation algorithms. (joint work with Valentine Genon-Catalot and Yves Rozenholc)

12h00 à 12h30 : **Jean-Baptiste AUBIN** (Univ. Paris 6).  
*Estimation fonctionnelle par projection adaptative.*

**Résumé.** ([résumé AUBIN](#))

12h30 à 13h00 : **Samuela LEONI-AUBIN** (Univ. Paris 6).  
*Sur la Vraisemblance Empirique dans des modèles à rapport de densités semi-paramétriques.*

**Résumé.** Nous considérons les problèmes d'estimation et de test à deux échantillons dans des modèles à rapport de densités semi-paramétriques. La vraisemblance empirique pose un problème d'irrégularité sous l'hypothèse nulle d'homogénéité. Nous montrons qu'une forme "duale" de la vraisemblance empirique est bien définie. Un test statistique, basé sur la forme duale de la vraisemblance empirique, est ensuite proposé. Les propriétés asymptotiques de la statistique du test sont étudiées sous l'hypothèse nulle et sous l'hypothèse alternative, et une approximation pour la fonction de puissance est déduite.

## Vendredi 20 janvier 2006

11h00 à 11h50 : **Christophe CROUX** (K.U. Leuven, Belgium).

*Granger causality analysis for business and consumers Surveys.*

**Résumé.** Each month business and consumer surveys are undertaken by the member states of the European Union. Consumers and companies have to provide their personal prediction about the future status of the consumption and production in Europe. They have to predict whether they expect an increase or decrease of these variables. Well-known indexes are derived from these surveys, such as the "consumer confidence index", or the "production expectation index". These surveys are costly and time-consuming, and their predictive power is therefore questioned. More precisely, we are, for example, interested to know whether these surveys about future production levels are able to make better forecasts than simple extrapolations based on the current and past production levels. This concept is known as Granger causality. We extend previous research in two ways, as we (i) explicitly allow for cross-country influences by using multivariate Granger-causality tests, and (ii) consider Granger-causality over Different planning horizons, by decomposing it over the spectrum.

12h00 à 12h50 : **Emma O'CONNOR** & **Nick FIELLER** (Sheffield University, U.K.)

*Statistical Testing of Medical Images.*

**Résumé.** Medical imaging provides a non-destructive method of direct investigation of effects of treatments on target tissues. This allows tissue to be examined on several occasions during the course of treatment, thus avoiding inter-individual variability. This presentation investigates methodology for assessing the statistical differences between images and hence the effectiveness (or otherwise) of the treatment. The particular study described here involved collection of three-dimensional images by MRI before and after treatment on individuals receiving one of a range of doses. Data extracted from the images were the separate voxel values of a parameter of interest. Statistical analysis focuses on the frequency distributions of these voxel values. The main analysis is based upon functional principal component analysis (FPCA) of the kernel density estimates obtained from each distribution. The analysis reveals interesting structure in the variation between the observed distributions with clear-cut and medically plausible interpretations of the first two components which together account for approximately ninety per cent of the variability between the empirical distributions. Statistical assessment of the effects of treatment is based upon the differences in PC scores between images before and after treatment; formal significance tests are provided by a randomization test.

## Vendredi 16 decembre 2005

11h00 à 11h50 : **Gilles CELEUX** (Université Paris-Sud). *Choix d'un modèle génératif de classification supervisée.*

**Résumé.** Le choix d'un modèle probabiliste pour l'analyse discriminante est l'objet de cet communication. Les critères classiques de sélection de modèle privilégient l'adéquation du modèle à la distribution jointe des variables explicatives et de la variable de groupe plutôt que la minimisation du taux d'erreur du classifieur associé. Nous proposons un nouveau critère, le Bayesian Entropy Criterion (BEC), qui permet de sélectionner un classifieur prenant en compte l'objectif décisionnel par la minimisation de l'entropie intégrée de classification. Il représente une alternative intéressante à la validation croisée qui est très coûteuse. Les propriétés asymptotiques du critère BEC sont présentées et des expériences numériques sur des données simulées et des données réelles montrent que ce critère a un comportement meilleur que BIC pour choisir le modèle minimisant l'erreur de classification et analogue à celui de la validation croisée.

12h00 à 12h50 : **Pierre VANDEKERKHOVE** (Université de Marne-la-Vallée). *Sur l'algorithme du bandit à deux bras dans un cadre ergodique.*

**Résumé.** ([resume\\_vandekerkhove](#))

## Vendredi 25 novembre 2005

11h00 à 11h50 : **Odile PONS** (INRA, Jouy-en-Josas). *Estimation de la loi de temps de séjour des processus de renouvellement Markoviens en dimension 2.*

**Résumé.** Self-consistency equations are established for the distribution functions of right and left censored of one and two dimensional variables and sojourn times of a Markov renewal process. They have a unique solution that equals the product-limit estimator if a hazard function may be defined.

12h00 à 12h50 : **Cristian PREDA** (Université Lille 2). *Regression models for functional data by reproducing kernel Hilbert spaces methods.* ([papier\\_preda](#))

**Résumé.** ([resume\\_preda](#)) Non-parametric regression models are developed when the predictor is a function-valued variable  $X = (X)_t \in T$ . Based on a representation of the regression function  $f(X)$  in a reproducing kernel Hilbert space such models generalize the classical setting used in statistical learning theory. Two applications corresponding to scalar and categorical response random variable are performed on stock-exchange and medical data. The results of different regression models are compared.

## Vendredi 21 octobre 2005

10h30 à 11h30 : **Ion GRAMA** (SABRES, Université de Bretagne-Sud). *Rate optimal adaptive estimation of the excess distribution function.*

**Résumé.** ([abstract](#))

11h30 à 12h30 : **Tiberiu SPIRCU** (Université "Carol Davila", Bucarest). *L'incertitude médicale, entre le traitement probabiliste et celui possibiliste.*

**Résumé.** L'évolution de la technique de l'information n'a pas influencé encore assez sérieusement la manière dont le médecin raisonne. Ses «opinions probables» sont encore subjectives, il travaille avec des informations incertaines, assez souvent inconsistantes. Les systèmes de support de la décision médicale, en évolution continue, nécessitent la quantification, d'une manière ou d'une autre, de l'incertitude médicale, aussi l'établissement de règles claires pour le traitement algorithmique de l'incertitude. Les règles classiques de la théorie des probabilités, particulièrement les deux théorèmes de Bayes, sont rarement utilisés dans la pratique médicale, parce que:

a) le nombre des calculs à effectuer pour obtenir une certaine probabilité est trop grand (même si irrelevant quand on utilise les ordinateurs);

b) les réseaux Bayesiens qui décrivent les connexions entre symptômes, syndromes et maladies sont d'une haute complexité;

c) on a de sérieuses difficultés à estimer les probabilités à priori, au moins pour les maladies rares.

Plusieurs modèles pour représenter les données incertaines ont été proposés:

1) les ensembles flous, introduits par Zadeh en 1965,

2) les crédibilités de Dempster-Shafer,

3) les possibilités de Dubois-Prade.

Dans ce travail, on essaye de combiner ces modèles.

## Vendredi 30 septembre 2005

10h30 à 11h30 : **Patrice BERTAIL** (CREST-LS, Paris). *Bootstrap regeneratifs pour les chaines de Markov.* ([PDF](#))

**Résumé.** Nous présentons une méthode de Bootstrap spécifique pour les chaînes de Markov Harris récurrentes positives, basée sur les propriétés de régénération et la technique de découpage introduite par Nummelin(1978) (et Atherya et Ney (1968)). Plus généralement, nous proposons une méthode pour découper une série temporelle en blocs qui sont pratiquement indépendants. L'idée principale sous-jacente à cette construction est :

1) de générer une séquence de temps de renouvellement approché pour la chaîne, à partir d'observation  $X_{\{1\}}, \dots, X_{\{n\}}$  et de la connaissance des paramètres d'une condition de minoration satisfaite par la densité de transition;

2) d'appliquer une variante de la méthodologie proposée par Datta et McCormick (1993) pour bootstrapper des fonctionnelles additives de chaînes de Markov atomiques.

Nous montrons que, dans le cas atomique, notre méthode hérite des propriétés au second ordre du bootstrap i.i.d, jusqu'à l'ordre  $O_{\{P\}}(n^{-1})\log(n)$  sous des conditions faibles. Dans le cas général, nous montrons la validité asymptotique de la procédure, sous l'hypothèse de l'existence d'un estimateur adéquat du noyau de transition, sous des conditions générales et obtenons la validité au second ordre de la méthode dans le cas stationnaire. Nous discutons des applications à des modèles spécifiques et présentons quelques résultats de simulations. Travail joint avec Stephan Clemencon.