

# Séminaire de statistique ENSAI

Année 2009-2010

Responsable Myriam Vimond

## Vendredi 25 juin 2010

Frédéric Lavancier, (Université de Nantes), *Résidus et tests d'adéquation pour les processus ponctuels de Gibbs marqués*.

**Résumé:** Les processus ponctuels spatiaux modélisent la répartition de points dans l'espace. Lorsqu'une marque est associée à chaque point, on parle de processus ponctuels marqués. Les processus ponctuels de Gibbs marqués en forment une très large classe de modèles et sont largement utilisés en pratique. Ils sont entièrement définis à l'aide d'une fonction d'interaction entre les points et d'une loi sur les marques. Des exemples seront présentés. Si on suppose une forme paramétrique pour la fonction d'interaction, différentes techniques permettent d'en estimer les paramètres (maximum de vraisemblance, pseudo-vraisemblance, etc). Etant données des observations, il est alors naturel de vouloir évaluer la qualité d'ajustement du modèle de Gibbs paramétrique choisi. La démarche classique consiste en l'étude des résidus de la modélisation. La notion de résidus pour les processus ponctuels a été récemment introduite par A.

Baddeley, R. Turner, J. Moller et M. Hazelton (cf [1]). Je montrerai le comportement asymptotique des résidus pour une large classe de modèles de Gibbs marqués. Ces résultats asymptotiques conduisent naturellement à la construction de tests d'adéquation. Je présenterai notamment une généralisation aux modèles de Gibbs du test des quadrants, utilisé pour tester le caractère poissonien d'une configuration de points. Ces résultats sont le fruit d'une collaboration avec J-F. Coeurjolly ([2]).

[1] A. Baddeley, R. Turner, J. Moller and M. Hazelton (2005) : « Residual Analysis for spatial point processes », JRSS B-67, pp617-666.

[2] J.-F. Coeurjolly and F. Lavancier : « Residuals and Goodness of fit tests for stationary marked gibbs point processes », arXiv:1002.0857.

## Vendredi 7 mai 2010

Christophe Demattei, (CHU de Nîmes), *Détection d'agrégats temporels et / ou spatiaux d'évènements ponctuels*

**Résumé:** La statistique de scan à fenêtres variables est le test de détection d'agrégats temporels ou spatiaux le plus utilisé. Elle consiste à scanner le domaine d'étude pour rechercher la zone qui maximise la vraisemblance. Le principal avantage de cette approche est sa puissance. L'un de ses inconvénients est de fixer a priori la forme paramétrique de l'agrégat potentiel. D'autres versions (elliptiques, non paramétriques) sont actuellement développées pour contourner ce manque de flexibilité.

Nous présentons ici une approche par régression sur données transformées, mise au point initialement dans le cas temporel, généralisée au cas spatial puis au cas spatio-temporel. L'idée de cette approche est de repérer les agrégats potentiels par des intervalles de temps dans lesquels le délai moyen séparant deux événements successifs est faible. La généralisation au cas spatial nécessite de définir une trajectoire afin de se ramener à la dimension 1. Cette trajectoire est définie en utilisant la distance d'un point à son plus proche voisin, compte tenu du trajet déjà effectué. La encore les agrégats potentiels sont repérés par des portions de trajectoire où la distance moyenne d'un point à son plus proche voisin est faible. Une correction pour effets de bord et inhomogénéité de la population à risque est effectuée en calculant l'espérance sous  $H_0$  de la distance d'un point à son plus proche voisin. Une p-valeur est obtenue pour chaque portion du modèle sélectionné par simulations de Monte Carlo. Enfin, la généralisation au cas spatio-temporel nécessite également la définition d'une distance spatio-temporelle. Cette distance est obtenue par une pondération des distances euclidiennes spatiale et temporelle, ce qui revient à dilater la fenêtre temporelle afin qu'elle ait la même étendue que le diamètre de la fenêtre spatiale.

Les différentes approches présentées sont illustrées sur différents jeux de données simulés ou réels issus du domaine de la santé (imagerie médicale, cancérologie), mais aussi de domaines tels que l'astrophysique ou la sismologie.

## **Vendredi 2 avril 2010**

Thomas Laloë, (Université Lyon I), *Sur Quelques Problèmes d'Apprentissage Supervisé et Non Supervisé*

**Résumé** : L'objectif de cette Thèse est d'apporter une contribution au problème de l'apprentissage statistique, notamment en développant des méthodes pour prendre en compte des données fonctionnelles. Dans la première partie, nous développons une approche de type plus proches voisins pour la régression fonctionnelle. Dans la deuxième, nous étudions les propriétés de la méthode de quantification dans des espaces de dimension infinie. Nous appliquons ensuite cette méthode pour réaliser une étude comportementale de bancs d'anchois. Enfin, la dernière partie est dédiée au problème de l'estimation des ensembles de niveaux de la fonction de régression dans un cadre multivarié.

## **Vendredi 15 janvier 2010**

Boubacar Maïnassara Yacouba, (Université de Lille 3), *Estimation des modèles VARMA structurels avec innovations linéaires non corrélées mais non indépendantes*

**Résumé** : Pour la modélisation des séries temporelles multivariées, les modèles VARMA (Vector AutoRegressive Moving-Average) occupent une place centrale. Ils sont généralement utilisés avec des hypothèses fortes sur le bruit qui en limitent la généralité. Dans ce travail, nous nous intéressons à l'analyse statistique de modèles vectoriels ARMA (VARMA) pour des processus qui peuvent avoir des dynamiques non linéaires très générales. Nous appelons VARMA forts les modèles standard dans lesquels le terme d'erreur est supposé être une suite iid, et nous parlons de modèles VARMA faibles quand les hypothèses sur le bruit sont moins restrictives. Dans un premier temps, nous étudions les propriétés asymptotiques du quasi-maximum de vraisemblance (QMLE) des paramètres d'un modèle VARMA sans faire l'hypothèse d'indépendance sur le bruit, contrairement à ce qui est fait habituellement pour l'inférence de ces modèles. Relâcher cette hypothèse permet aux modèles VARMA faibles de couvrir une large classe de processus non linéaires. Nous faisons

des hypothèses d'ergodicité et de mélange afin d'établir la convergence forte et la normalité asymptotique de l'estimateur du QMLE. Ensuite, nous accordons une attention particulière à l'estimation de la matrice de variance asymptotique qui a la forme "sandwich"  $\Omega := J^{-1} I J^{-1}$ , et qui peut être très différente de la variance asymptotique standard dont la forme est  $\Omega := 2J^{-1}$ . Nous établissons la convergence d'un estimateur de  $\Omega$ . Enfin, des versions modifiées des tests de Wald, du multiplicateur de Lagrange et du rapport de vraisemblance sont proposées pour tester des restrictions linéaires sur les paramètres libres du modèle.

## **Vendredi 4 décembre 2009**

[Gilbert Mackenzie](#), (University de Limerick), *Advances in Multivariate Survival Modelling and h-likelihood methods of estimation*

**Résumé** : In this talk we will concentrate mainly on non-PH survival models (MacKenzie, 1996, 1997) and on h-likelihood as a method of inference, areas of survival modelling, which in the last decade have not received all of the attention which they deserved. The reasons are many-faceted, but basically they have been over-shadowed by the the march of the martingale machinery applied to Cox's (1972) PH model. We shall not of course forget to mention PH models because, *emph{inter alia}*, h-likelihood methods apply there too (Ha, Lee & MacKenzie, 2007)

First we trace the rationale and development of the generalized time-dependent logistic (GTDL) family of survival models defined by  $\lambda(t; x) = \lambda_0 \exp(t\alpha + x'\beta) / [1 + \exp(t\alpha + x'\beta)]$ , where:  $\lambda_0 > 0$ ,  $\alpha$  (real) are scalars and  $\beta$  is vector of (p+1) regression parameters measuring the influence of the covariates  $x'$ . Of the three basic models in the family, we focus on one non-PH regression model. The genesis of this model is of interest and we indicate several derivations including a frailty interpretation, its relationship to Fisher's Z distribution and the connection with Aalen's (1988) class. Next we consider an identifiability problem which necessitates extensions to the model via a Bayesian approach (Louzada-Neto, Cremasco & MacKenzie, 2009), via univariate frailty methods with structural dispersion (MacKenzie & Lynch, 2007) and via multivariate survival modelling methods (Ha & MacKenzie, 2009). In this section we will also note a multivariate version of GTDL model (Blagojevic & MacKenzie, 2004).

The use of h-likelihood methods of estimation are gaining in popularity and we describe the idea of Extended Restricted Likelihood (ERL) underpinning the approach of Lee & Nelder (1996, 2001), illustrating its use in a comparison of non-PH and PH models in the multivariate survival setting. A key advantage of the method is that it obviates the need to integrate out the random effects.

However, compared with classical parametric models, model selection in semi-parametric frailty models is complicated by the presence of a nonparametric (baseline hazard) nuisance function,  $\lambda_0(t)$ , the dimension of which is an increasing function of the sample size  $n$ . Accordingly, we indicate how extensions of two HGLM model selection criteria to frailty models can be developed as AICs, via a profile likelihood, after eliminating the high-dimensional (fixed) nuisance parameters in  $\lambda_0(t)$ .

## **Vendredi 13 novembre 2009**

[Paul Rochet](#), (Université de Toulouse), *Maximum entropy method applied to survey sampling*

**Résumé** : Calibration methods have become increasingly studied in survey sampling over the last decades. By viewing calibration as an inverse problem, this article extends the calibration technique

in the presence of complete auxiliary information by using a maximum entropy method (MEM). Finding the optimal weights is achieved by considering random weights and looking for a discrete distribution which maximizes an entropy under the calibration constraint. This method enables to incorporate prior informations to the problem, giving a Bayesian interpretation to arbitrary settings existing in calibration. Asymptotic properties of calibrated estimators are studied in a general framework by applying some other predominant methods in survey sampling like generalized calibration or instruments technique. We point out the relation between calibration and linear regression, and extend it to non-linear estimation. Optimality results are obtained using a generalization of the maximum entropy method, yielding a better asymptotic efficiency than classical calibration.

### **Vendredi 16 octobre 2009**

[Olivier Lopez](#), (Université Paris VI), *Tests d'adéquation à des modèles de régression paramétriques en présence de censure*

**Résumé** : Nous proposons de nouvelles procédures de tests non paramétriques d'adéquation pour des modèles de régression où la variable expliquée est censurée aléatoirement à droite. Nous étudions d'abord le cas où la censure est indépendante des variables impliquées dans le modèle de régression, puis le cas où la censure peut dépendre des variables explicatives. Le cadre général dans lequel s'appliquent nos résultats permet d'englober notamment le cas des modèles de régression paramétriques portant sur l'espérance conditionnelle ou les quantiles conditionnels.

### **Vendredi 18 septembre 2009**

[David Degras](#), (The University of Chicago), *Simultaneous confidence bands for nonparametric regression with repeated measurements data*

**Résumé** : We look into nonparametric regression with repeated measurements collected on a fine grid. An asymptotic normality result is obtained in a function space. This result can be used to build simultaneous confidence bands (SCB) for various tasks in statistical exploration, estimation and inference. Two applications are proposed: one is a SCB procedure for the regression function and the other is a goodness-of-fit test for linear regression models. The first one improves upon other available methods in terms of accuracy while the second can detect local departures from a parametric shape, as opposed to the usual goodness-of-fit tests which only track global departures. A numerical study is also provided.