Groupe de travail STAPH : Statistique Fonctionnelle et Opératorielle

Recueil de résumés 2004-2005

Coordinateurs
A. Boudou, H. Cardot, F. Ferraty, Y. Romain,
P. Sarda, P. Vieu et S. Viguier-Pla

Résumé

Ce document a pour objectif de présenter les résumés (plus ou moins détaillés selon les souhaits de leurs auteurs) des divers exposés qui ont eu lieu lors des séances du groupe de travail STAPH durant l'année universitaire 2004-2005.

Rappelons que ce groupe de travail en Statistique Fonctionnelle et Opératorielle, créé il y a quelques années au sein du Laboratoire de Statistique et Probabilités de Toulouse, s'inscrit dans la dynamique actuelle autour des divers aspects fonctionnels de la statistique moderne. Les exposés qui sont présentés traitent de divers aspects de la Statistique Fonctionnelle (estimation nonparamétrique, statistique opératorielle, modèles de réduction de dimension, modèles pour variables fonctionnelles, ...) ; ils sont de nature différentes (exposés didactiques ou bibliographiques, exposés de résultats nouveaux en Statistique Appliquée et/ou Théorique, ...) ; ils témoignent enfin de la perpétuelle ouverture de la démarche par la grande diversité des exposants.

Pour terminer signalons que l'intégralité des activités de ce groupe de travail est disponible sur notre page web :

$$http://www.lsp.ups-tlse.fr/staph.html$$

Abstract

We present the abstracts (of size more or less important according to the wishes of their authors) of the several different talks given during the sessions of the working group STAPH along the academic year 2004-2005. This group in Functional and Operatorial Statistics is born a few years ago at the Laboratoire de Statistique et Probabilités of the Université Paul Sabatier de Toulouse, and its aim was to participate at the actual dynamic existing around the different functional features of modern statistics.

These talks were about different functional topics (nonparametric estimation, statistics of operators, models for functional data, models for dimension reduction, ...). They were of different kinds (didactic, bibliografic, applied, theoric, ...) and were presented by a large variety of statisticians.

As a conclusion, note that all the activities of this group are reachable through the following web adress :

$$http://www.lsp.ups-tlse.fr/staph.html.$$

Sumario

Este documento presenta resumens (mas o menos cortos segun los deseos de sus autores) de charlas que han sido presentadas durante las sesiones de trabajo del grupo STAPH durante el ano academico 2004-2005. Este grupo de trabajo en el campo de Estadistica Funcional y Operatorial ha sido creado hace algunos anos en el Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier de Toulouse, para animar investigaciones en varios aspectos funcionales de la estdistica moderna.

Estas conferencias fueran sobre temas variados (estimacion noparametrica, estadistica de operatores, modelos para variables funcionales, modelos de reduccion de dimension, ...) y fueran de tipos differentes (conferencias didacticas o bibliograficas, presentacion de resultados nuevos en estadistica teorica o/y applicada, ...).

Al final, queremos apuntar que todas nuestras actividades pueden ser consultadas a la direccion :

$$http://www.lsp.ups-tlse.fr/staph.html.$$

Modèles Non-paramétriques et/ou Variables Fonctionnelles

Frédéric FERRATY et **Philippe VIEU**

Adresse pour correspondance :
Laboratoire de Statistique et Probabilités,
Université Paul Sabatier, 31062 Toulouse Cedex 4
e-mail : ferraty@cict.fr, vieu@cict.fr

Séance du 11 Octobre 2004

Résumé

L'objectif est de faire un rapide bilan de l'état de l'art en matière de *statistique non-paramétrique* pour *variables fonctionnelles*. Après une rapide présentation de ce cadre de *double dimension infinie*, nous ferons une présentation rapide de quelques résultats récents et l'accent sera mis sur plusieurs problèmes ouverts.

La première partie sera consacrée à la présentation de définitions précises de nature à dissiper les ambiguïtés/malentendus qui existent autour des mots *non-paramétrique* et *fonctionnel*. Qu'est ce qu'un modèle paramétrique (resp. non-paramétrique) pour variables réelles (resp. variables fonctionnelles) ? Que signifie l'expression courante *estimation fonctionnelle* ? Nous en viendrons alors à la définition de *modèle non-paramétrique fonctionnel*, et nous expliquerons comment les problèmes classiques d'estimation fonctionnelle se traduisent naturellement en problèmes d'*estimation opératorielle*.

Dans un deuxième temps, nous présenterons quelques résultats asymptotiques récents qui sont pour la plupart issus de Ferraty-Vieu (2004). Les propriétés asymptotiques de divers estimateurs non-paramétriques fonctionnels, correspondant à des problèmes d'estimation opératorielle différents, seront rapidement présentées. Nous insisterons plus particulièrement sur le lien existant entre les *vitesses de convergence*, la *double dimension infinie* et les questions de *probabilités de petites boules* dans des espaces de dimension infinie.

Dans sa forme, cet exposé accordera une place particulièrement importante aux problèmes/modèles de régression ainsi qu'à la mise en évidence de la multi-

tude de problèmes ouverts par ce nouveau champ de la statistique.

Références

Ferraty, F. and Vieu, P. (2004). Nonparametric moels for functional data, with application in regression, time series prediction and curve discrimination. *J. Nonparametric Statistics*, **16** 111-127.

Contributions à la modélisation statistique fonctionnelle

Hervé CARDOT

INRA Toulouse, Mathématiques Informatique et Applications & Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse.
e-mail : cardot@toulouse.inra.fr

Habilitation à Diriger des Recherches
soutenue à l'Université Paul Sabatier

le 20 Octobre 2004

Résumé

La présentation débutera par un exposé de mes travaux sur les modèles pour données fonctionnelles (régression linéaire, modèles linéaires généralisés et quantiles conditionnels). J'expliquerai en quoi l'estimation d'un modèle de régression dans un cadre fonctionnel est généralement un *problème mal posé*. Il devient alors nécessaire d'introduire une procédure de *régularisation* (par une pénalisation ou une réduction de la dimension) afin d'exhiber un estimateur convergent. Les résultats de convergence sont obtenus sous des hypothèses faibles et je discuterai comment il est possible de les améliorer.

Je détaillerai ensuite deux exemples concrets portant sur l'utilisation de chroniques d'images satellite basse résolution en agriculture qui illustrent bien à la fois le métier de *statisticien modélisateur* à l'INRA et le potentiel des méthodes nonparamétriques et fonctionnelles. Il s'agit de problèmes de *désagrégation* d'une information spatiale à l'aide d'observations temporelles. La première question concerne l'estimation du plan d'occupation des sols à l'aide d'une chronique d'observations à la résolution kilométrique (suite temporelle de pixels basse résolution). Nous verrons pourquoi des approches naturelles basées sur une description physique du phénomène d'agrégation se montrent inefficaces par rapport à un modèle de type multilogit fonctionnel.

La deuxième question concerne l'estimation des réponses locales des cultures. Ces réponses dépendent de variables telles que la nature du sol, les pratiques des agriculteurs, ..., qui ne sont pas toutes observables. C'est pourquoi nous proposons un modèle nonparamétrique à *effets aléatoires* de désagrégation qui permet de

tenir compte des variations locales des réponses et des mécanismes individuels. D'un point de vue "données fonctionnelles", ce modèle est une généralisation de l'ACP lorsqu'on observe des mélanges bruités de fonctions aléatoires issues de différentes populations.

Un test d'adéquation global de la fonction de répartition conditionnelle

Sandie FERRIGNO

Adresse pour correspondance :
Laboratoire de Statistique et Probabilités,
Université Paul Sabatier, 31062 Toulouse Cedex 4
e-mail : ferrigno@math.univ-montp2.fr

Séance du 25 Octobre 2004

Résumé

Soient X et Y, deux variables aléatoires. De nombreuses procédures statistiques permettent d'ajuster un modèle à ces données dans le but d'expliquer Y à partir de X. La mise en place d'un tel modèle fait généralement appel à diverses hypothèses que l'on doit valider pour justifier son utilisation.

Dans ce travail, on propose une approche globale où toutes les hypothèses faites pour asseoir ce modèle sont testées simultanément. Plus précisément, on construit un test basé sur une quantité qui permet de canaliser toute l'information liant X à Y : la fonction de répartition conditionnelle de Y sachant $\{X = x\}$ définie par $F(y|x) = P(Y \leq y|X = x)$. Notre test compare la valeur prise par l'estimateur polynômial local de $F(y|x)$ à une estimation paramétrique du modèle supposé et rejette sa validité si la "distance" entre ces deux quantités est trop grande. Dans un premier temps, on considère le cas où la fonction de répartition supposée est entièrement spécifiée et, dans ce contexte, on établit le comportement asymptotique du test.

Dans une deuxième partie, on généralise ce résultat au cas plus pratique où le modèle supposé contient un certain nombre de paramètres inconnus. On étudie ensuite la puissance locale du test en déterminant son comportement asymptotique local sous des suites d'hypothèses contigües.

Enfin, on propose un critère de choix de la fenêtre d'ajustement qui intervient lors de l'étape d'estimation polynômiale locale de la fonction de répartition

conditionnelle.

Références

[1] Alcalá, J.T., Cristóbal, J.A. & González-Manteiga, W. (1999). Goodness-of-fit tests for linear models based on local polynomials. *Statistics & Probability Letters*, **42**, 39-46.

[2] Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

[3] Härdle,W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**(4), 1926-1947.

Test de sélection de distribution dans une classe de modèles surdispersés

Célestin C. KOKONENDJI

Adresse pour correspondance :
Université de Pau et des Pays de l'Adour
Laboratoire de Mathématiques Appliquées
Département STID - IUT des Pays de l'Adour
Avenue de l'Université - 64000 Pau, France.
e-mail : celestin.kokonendji@univ-pau.fr

Séance du 15 Novembre 2004

Résumé

Dans cet exposé, nous nous intéresserons particulièrement à une famille des modèles surdispersés dite de Hinde-Demétrio, laquelle contient les deux modèles bien connus : binomial négatif et arcsinus strict (Letac & Mora, 1990 ; Kokonendji & Khoudar, 2004). Il s'agit d'une nouvelle famille de lois de probabilité à trois paramètres, où le troisième paramètre permet de discriminer le modèle exponentiel de dispersion associé (Jørgensen, 1997). Cette famille Hinde-Demétrio est complémentaire à des familles de la littérature (e.g., Hougaard *et al.*, 1997 ; Whalin & Paris, 2002 ; Castillo & Pérez-Casany, 1998, 2004).

Dans Kokonendji *et al.* (2004), nous avons complété le travail de Hougaard *et al.* (1997) puis nous avons décrit entièrement la nouvelle classe de lois dite de Hinde-Demétrio. En effet, pour améliorer significativement l'ajustement de certaines données, Hougaard *et al.* (1997) ont considéré une grande famille de lois de mélange de Poisson avec des lois stables-positives, qui sont des modèles exponentiel de dispersion de fonctions variance puissances-positives (e.g., Tweedie, 1984 ; Jørgensen, 1997). Nous avons donc appelé une fermeture de cette famille de lois discrètes la classe *Poisson-Tweedie*, que nous avons caractérisée par leur fonction variance unité :

$$V_p^{PT}(v) = v + v^p \exp\{(2-p)\Psi_p(v)\}, \quad v > 0, p \geq 1,$$

où $\Psi_p(v)$ est la fonction inverse (généralement implicite) de la dérivée première de la fonction cumulante. Les densités des lois associées à la classe Poisson-Tweedie

sont exhibées pour tout $p \geq 1$, et cette classe contient le modèle binomial négatif avec $p = 2$ et le modèle Poisson-inverse-gaussien avec $p = 3$ (Holla, 1966 ; Sichel, 1971).

Comme une approximation (en termes de fonction variance unité), nous avons introduit puis décrit complètement la classe *Hinde-Demétrio*, qui est caractérisée par leur "simple" fonction variance unité (e.g. Hinde & Demétrio, 1998, page 14) :

$$V_p^{HD}(v) = v + v^p, \quad p \in \{0\} \cup [1, +\infty[,$$

où $v > -1$ pour $p = 0$ et $v > 0$ pour $p \geq 1$. Autrement dit, étant donné $p \in \{0\} \cup [1, +\infty[$, la fonction de probabilité individuelle d'une loi $\mathcal{HD}_p(\theta, \sigma)$ de Hinde-Demétrio s'écrit :

$$P(x; p; \theta, \sigma) = c(x; p; \sigma) \exp\{\theta x - \sigma k_p(\theta)\}, \quad x \in S_p,$$

où $\theta \in \Theta_p \subseteq \mathbb{R}$ est le paramètre canonique, $\sigma > 0$ est le paramètre de Jørgensen ou de puissance de convolution, $c(x; p; \sigma)$ est la constante de normalisation, $k_p(\theta)$ est la fonction cumulante vérifiant $k_p''(\theta) = V_p^{HD}(k_p'(\theta)) = k_p'(\theta) + (k_p'(\theta))^p$, et le support S_p est tels que $S_0 = \{-1\} \cup \mathbb{N}$, $S_1 = 2\mathbb{N}$ et $S_p = \mathbb{N} + p\mathbb{N}$ pour $p > 1$. Des cas particuliers sont donnés par une translatée positive de Poisson pour $p = 0$, une multiple de Poisson pour $p = 1$, la binomiale négative pour $p = 2$, et l'arcsinus stricte pour $p = 3$ (Kokonendji & Khoudar, 2004). Quand $p \neq 0, 1, 2, 3$, la densité $P(x; p; \theta, \sigma)$ n'a pas une forme explicite ou facilement exploitable, même si leur fonction cumulante peut se mettre sous la forme :

$$\begin{aligned} k_p(s) &= \sum_{k=0}^{\infty} \frac{\Gamma[k + 1/(p-1)]}{k!\Gamma[1/(p-1)]} \frac{\exp\{s[1 + k(p-1]\}}{1 + k(p-1)} \\ &= e^s \, {}_2F_1\left(\frac{1}{p-1}, \frac{1}{p-1}; \frac{p}{p-1}; e^{s(p-1)}\right), \quad s < 0, \quad p > 1, \end{aligned}$$

où ${}_2F_1\,(a, b; c; z) = 1 + \frac{ab}{c}\frac{z}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2!} + \cdots$ est la fonction hypergéométrique gaussienne (e.g. Rainville, 1960). Nous avons terminé le papier Kokonendji *et al.* (2004) par une tentative d'estimation de p par la méthode des moments afin de sélectionner le modèle adéquat dans ces deux classes.

Dans le travail référencé par Kokonendji & Malouche (2004) et dont il est question dans cet exposé, nous nous sommes intéréssés aux modèles de Hinde-Demétrio concentrés sur $\mathbb{N}$ pour l'analyse statistique des données de comptage surdispersées ; i.e., nous avons considéré $p \in \{2, 3, \cdots\}$ (e.g. Cameron & Trivedi, 1986). Tout d'abord, en utilisant une interprétation probabiliste de loi $\mathcal{HD}_p(\theta, \sigma)$ de Hinde-Demétrio, nous montrerons la propriété suivante : si $r_x = x\,P(x; p; \theta, \sigma)/P(x-1; p; \theta, \sigma) = r_x(p; \theta, \sigma)$ pour tout $\theta < 0$ et $\sigma > 0$ alors

$$r_1 = r_2 = \cdots = r_{p-1} < r_p, \quad \forall p \in \{2, 3, \cdots\};$$

de plus, nous aurons $r_{p+1} \neq r_p$ et $r_{p+1} > r_1$. A la suite, nous proposerons de tests statistiques sur p, d'hypothèses :

$$H_{0p}^{ij} : r_i = r_j \quad \text{contre} \quad H_{1p}^{ij} : r_i < r_j \quad (1 \leq i < j \leq p),$$

afin de sélectionner le modèle approprié dans la famille de Hinde-Demétrio pour un jeu de données surdispersées. Enfin, nous concluerons en illustrant la technique proposée sur des données simulées ainsi que sur les données réelles de la littérature (e.g. Greenwood & Yule, 1920 ; Kokonendji & Khoudar, 2004). Nous discuterons également du cas limite $p = 2$, correspondant au modèle prototype binomial négatif, lequel se trouve souvent à l'intersection de plusieurs classes de modèles surdispersés par rapport au modèle poissonnien.

Références

Cameron, A. C. & Trivedi, P. K. (1986). Econometric models based on count data : comparisons and applications of some estimators and tests. *J. Appl. Econometrics* **1**, 29-53.

Castillo, J. & Pérez-Casany, M. (1998), Weighted poisson distribution for overdispersion and underdispersion situations, *Ann. Inst. Statist. Math.* **50**, 567-585.

Castillo, J. & Pérez-Casany, M. (2004). Overdispersed and underdispersed Poisson generalizations. *J. Statist. Plann. Inference*, to appear.

Greenwood, M. & Yule, G.U. (1920), An inquiry into the nature of frequency distributions representative of multiple happenings with particular referee to the occurrence of multiple attacks of disease or of repeated accidents, *J. R. Statist. Soc. Ser. A* **83**, 255-279.

Hinde, J. & Demétrio, C.G.B. (1998). *Overdispersion : Models and Estimation.* São Paulo : ABE.

Holla, M.S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika* **11**, 115-121.

Hougaard, P., Lee, M-L.T. & Whitmore, G.A. (1997). Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes, *Biometrics* 53, 1225-1238.

Jørgensen, B. (1997). *The theory of dispersion models,* Chapman & Hall, London.

Kokonendji, C.C. & Khoudar, M. (2004). On strict arcsine distribution. *Communications in Statistics - Theory and Methods* **33** (5), 993-1006.

Kokonendji C.C., Demétrio C.G.B. & Dossou-Gbété S. (2004). Some discrete exponential dispersion models : Poisson-Tweedie and Hinde-Demério classes. *SORT* (*Statistics and Operations Research Transactions*) **28** (2) (à paraître).

Kokonendji, C.C. & Malouche, D. (2004). Selecting test of distribution in the Hinde-Demétrio family. *The Annals of Statistics* (soumis).

Letac, G. & Mora, M. (1990), Natural real exponential families with cubic variance functions, *Ann. Statist.* **18**, 1-37.

Rainville, E. D. (1960). *Special Functions*. Chelsea, New York.

Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. *Proceeding of the Third Symposium on Mathematical Statistics,* N.F. Laubscher (editor), 51-97. C.S.I.R., Pretoria.

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

Whalin, J. F. and Paris, J. (2002). A general family of overdispersed probability laws. *Belgian Actuar. Bull.* **2** 1-8.

Méthodes statistiques pour l'évaluation du risque lié à la présence de contaminants dans les aliments

Amélie CRÉPET et **Jessica TRESSOU**

Adresse pour correspondance :
INRA-INAPG, Unité Mét@Risk
16 rue Claude Bernard
75231 PARIS cedex 05
crepet@inapg.fr, jessica.tressou@inapg.inra.fr

Séance du 19 Novembre 2004
Exceptionnellement couplée avec le
séminaire de l'Unite BIA de l'INRA Toulouse

Résumé

Plusieurs outils statistiques ont été développés pour l'évaluation du risque alimentaire dans le cadre de ce travail. Le paramètre que l'on cherche à estimer à partir de données de consommation et contamination est la probabilité de dépasser une dose tolérable par l'organisme, probabilité que nous appelons "risque". Dans un premier temps, nous avons développé une méthode basée sur la théorie des valeurs extrêmes permettant de quantifier des "risques" faibles Afin de quantifier précisément un risque plus important provenant d'aliments multiples, une méthode de simulation utilisant des arguments de U-Statistiques, des techniques de Jackknife et de Bootstrap a été développée. De plus, un autre travail est en cours pour intégrer au modèle la censure des données de contamination. On présentera ces travaux, sans trop entrer dans les détails techniques, pour vous familiariser avec le problème. Enfin, on présentera les éléments à disposition pour caractériser les populations à risque afin d'en discuter.

Le modèle de régression non paramétrique pour données fonctionnelles : développements récents

André MAS

Adresse pour correspondance :
Equipe de Probabilités Statistiques, UMR CNRS 5149, CC051, Université Montpellier 2, place Eugène Bataillon, 34095 Montpellier Cedex
e-mail : mas@math.univ-montp2.fr

Séance du 6 décembre 2004

Résumé

Le modéle de régression non paramétrique pour données fonctionnelles généralise la classique régression non linéaire aux données infini-dimensionnelles. Il est de la forme :

$$y_i = r\left(X_i\right) + \varepsilon_i$$

où les y_i sont des observations scalaires, les X_i sont des éléments aléatoires à valeur dans une espace de Banach séparable E et où r est donc un fonction de E vers $\mathbb{R}$ (on supposera les ε_i indépendants des X_i). Frédéric Ferraty et Philippe Vieu ont proposé d'estimer r en s'inspirant de la méthode du noyau et des résultats de convergence ont déjà été obtenus.

Les premiers résultats qui seront exposés ont trait au calcul des constantes exactes dans les développements asymptotiques du biais et de la variance de notre estimateur (pour la convergence ponctuelle). Ces constantes étaient jusqu'alors inconnues, leur expression demeure assez théorique et donne lieu lors de simulations à une mise en oeuvre par bootstrap. Nous énonçons comme Corollaire un TCL pour l'estimateur.

Une des problématiques originales du modèle est liée à l'absence de densité pour des X_i en dimension infinie. Celle-ci (et ses dérivées) apparaît pourtant dans l'expression du biais ou de la variance dans le cas multidimensionnel. La densité est ici remplacée par une probabilité de petites boules, qui peut être facilement estimée mais qui complique considérablement les calculs. Jusqu'à présent les auteurs effectuaient des hypothèses fortes sur le comportement de ces probabilités de petites boules de type "existence d'un développement fractal". Nous énonçons

nos résultats dans un cadre plus général. Mais cette approche, même si elle permet d'englober une classe plus large de processus X, soulève de nombreuses questions.

Références

Ferraty, F., Mas, A. and Vieu, Ph. (2004). Recent advances on nonparametric functional regression. Preprint.

18

.

Quelques réflexions sur l'opératoriel.

Yves ROMAIN

Adresse pour correspondance :
Laboratoire de Statistique et Probabilités
Université Paul Sabatier, Toulouse
e-mail : romain@cict.fr

Séance du 13 décembre 2004

Résumé

Á partir de la lecture d'un article de Peter Atkins (un chapitre d'un ouvrage général, "Le doigt de Galilée"), on part dans des contrées bien éloignées de nos préoccupations habituelles. L'arithmétique est une discipline peu abordée par le statisticien et a priori son intersection avec la statistique nous semble quasiment vide. On y rencontre, par exemple, la conjecture de Goldbach, l'Hotel de Hilbert, une machine de Turing universelle, un théorème de Gödel ou encore l'hypothèse du continu...
Cette dernière nous permet quand même d'entrevoir des préoccupations communes et notamment dès lors qu'on s'interesse à la Statistique de fonctions aléatoires ou de données de courbes.
De son coté, la dimension infinie a souvent servi de frontière pour désigner la branche maintenant dénommée " Statistique opératorielle". Paradoxalement, l' "opératoriel" a aussi une signification propre en dimension finie : il s'agit de pratiquer la Statistique "intrinsèquement" c'est à dire sans utilisation de bases ni du langage matriciel. Divers avantages de considérer la Statistique par les opérateurs sont brièvement revus et commentés.
Cette présentation succincte suscitera naturellement une discussion au sein du GT sur les domaines et/ou mots clés de nos travaux passés et à venir.

20

.

Parallélisation de programmes et statistique fonctionnelle :

Table ronde animée par

Sébastien DEJEAN(1) et **Nicolas RENON**(2)

Adresses pour correspondance :
(1) LSP, Unviv. Paul Sabatier, Toulouse
(2) CICT, Unviv. Paul Sabatier, Toulouse
e-mail : sebastien.dejean@math.ups-tlse.fr and renon@cict.fr

Séance du 24 Janvier 2005

Résumé

L'objectif de cette séance est d'ouvrir une discussion sur l'utilisation de méthodes de calcul intensif pour la statistique fonctionnelle.
Dans un premier temps, nous présenterons très succintement, le principe de la programmation parallèle, puis les possibilités pratiques offertes par le CICT pour mettre en oeuvre ces méthodes au travers de Calmip.
Ensuite, nous présenterons des travaux en cours au sein du LSP sur la programmation parallèle. Nous aborderons d'abord les gains obtenus au niveau des temps de calcul qui ont permis de repousser les limites pratiques rencontrées sur la machine Ondine en programmation séquentielle. Ensuite, nous évoquerons l'investissement initial nécessaire pour assurer le portage de programmes séquentiels existants vers des programmes parallèles.
La discussion sera ensuite ouverte sur les possibilités d'initier des travaux de programmation parallèle sur des thèmes en relation avec la statistique fonctionnelle.

Quelques adresses utiles

Le site de Calmip : *http ://www.calmip.cict.fr/spip/*
L'adresse e-mail pour les demandes : *admin.calmip@cict.fr*
Le site de de l'IDRIS pour les cours sur le C, le FORTRAN, OpenMP, MPI... : *http ://www.idris.fr/docs/docu/support_cours/*

.

Estimation de quantiles conditionnels avec des arbres de régression : intervalles de prédiction du rendement de maïs

Jacques-Eric BERGEZ, Hervé CARDOT* et Frédérick GARCIA

Adresse pour correspondance :
LSP & INRA Toulouse,
Unité Biométrie et Intelligence Artificielle
31326 Castanet-Tolosan
cardot@toulouse.inra.fr

Séance du 7 Février 2005

Résumé

L'objectif de ce travail est l'obtention, en cours de saison, d'informations sur le rendement du maïs et la détermination des stades et des facteurs importants.

Pour cela nous disposons d'un modèle de croissance "mécaniste" du maïs (MODERATO) alimenté par des variables climatiques observées quotidiennement (température, pluviométrie, ETP, ...) et des conduites de culture issues de stratégies "optimales". Les variables climatiques sont simulées par un générateur de climat calibré à partir des observations climatiques à Blagnac ces 40 dernières années.

Nous disposons *in fine* d'une quinzaine de variables (climat, évolution biophysiques, décisions de l'agriculteur) observées quotidiennement et il s'agit d'utiliser au mieux cette information pour en déduire un intervalle de prédiction du rendement avant la récolte.

Cette question est abordée à l'aide des quantiles (conditionnels) qui permettent de fournir des intervalles de prédiction pour chaque horizon. Des phénomènes non linéaires et complexes entrent en jeu et une approche nonparamétrique semble la plus appropriée.

Nous montrons comment les arbres de régression, initialement construits pour estimer l'espérance conditionnelle, peuvent être adaptés au cas des quantiles conditionnels.

Les variables sont très nombreuses et pour certaines fonctionnelles. Différentes stratégies de réduction de la dimension sont comparées (ACP fonctionnelle, construction d'indices, ...).

Nous constatons sur cet exemple que la connaissance précise du climat n'est pas primordiale si l'on dispose de certaines variables biophysiques (faciles à mesurer), telle que "matière séche aérienne". Par ailleurs les mesures effectuées les 100 premiers jours après le semis ne sont pas informatives.

Références

J.E. Bergez, H. Cardot and F. Garcia (2005). Quantile regression trees for yield prediction using a simulation model. *Preprint*.

Bergez, J.-E., Debaeke, Ph., Deumier, J.-M., Lacroix, B., Leenhardt, D., Leroy, P., Wallach, D., 2001. MODERATO : an object-ori ented decision model to help on irrigation scheduling for corn cro p. *Ecological Modelling*, 137(1) : p43-60.

Bergez, J.E. and Garcia, F., 2002. A hierarchical partitioning method for optimizing irrigation strategies. *Proc. of the First European Workshop on Sequential Decisions under Uncertainty in Agriculture and Natural Resources*, Toulouse, France.

Breiman, L., Friedman, J.H., Olshen, R. and C.J. Stone, 1984 . *Classification and Regression Trees*. Chapman & Hall.

Cardot, H., Snoeck, L., Bergez, J.E. and Garcia, F., 2002. Quantile regression trees for predicting irrigated maize crop yields. *Proc. of the First European Workshop on Sequential Decisions under Uncertainty in Agriculture and Natural Resources*, Toulouse, France.

Chaudhuri, P. and Loh, W.Y, 2002. Nonparametric estimation of conditional quantiles using regression trees. *Bernouilli* , 8 :561-576.

Koenker, R. and Basset, G.S., 1978. Quantile regression. *Econometrica*, 46 :33-50.

Racsko, P., Szeidl, L. and Semenov, M., 1991. A serial approach to local stochastic weather models. *Ecological Modelling*, 57.

Change point detection and extremes of random processes

Daniela JARUVSKOVA

Adresse pour correspondance :
Czech Technical University, Dept. of Mathematics
Thákurova 7, CZ – 166 29 Praha 6, Czech Republic
jarus@mat.fsv.cvut.cz

Séance du 11 Février 2005

Exceptionnellement couplée avec le
séminaire de l'équipe SMASH de l'Univeristé de Toulouse 2

Abstract

Testing for a change in the mean μ of a sequence of independent normally distributed random variables $Y_1, \ldots, Y_n$ may be based on the maximum type test statistic. For a large number of observations n the limit behavior of this statistic (under the null hypothesis of no change) is given by a maximum of a non-differentiable Gaussian process. However, there are many situation when the change may occur in more parameters at the same time. The change in both parametrs μ and σ^2 as well as appearance of linear trend may serve as examples. Here, the limit behavior of the maximum type test statistic is given be the behavior of an extreme of a χ^2 process. The χ^2 process is defined as a square distance of a multivariate Gaussian process from the origin. Its components may be of the same type (i.e. all differentiable or all non-differentiable) or of different types. Moreover, they can be independent as well as dependent. The quite general theory of extremes of χ^2 processes on a fix interval was obtained by Piterbarg (1988) and on an increasing interval by Albin (1990). These results may be succesfully applied for studying the limit behavior of test statistics for change point detection, see e.g. Albin, Jarušková (2003).

References

Albin J.M.P., Jaršková D. : On a test statistic for linear trend, Extremes 6, 247 – 258, 2003.

Albin J.M.P. : On extremal theory for stationary processes, The Annals of Probability 18, 92 – 128, 1990.

Piterbarg V.I. : Asymptotic Methods in the theory of Gaussian processes and fields, The American Mathematical Society, ISBN 0-8218-0423-5, (translated from Russian.)

Tests for continuity of regression function

Jaromir ANTOCH *, Gérard GRÉGOIRE, Marie HUVSKOVA

Adresse pour correspondance :
Charles University of Prague, Faculty of Mathematics and Physics,
Dept. of Statistics, Sokolovská 83, CZ – 186 75 Praha 8,
Czech Republic
jaromir.antoch@mff.cuni.cz

Séance du 11 Février 2005

Exceptionnellement couplée avec le
séminaire de l'équipe SMASH de l'Univeristé de Toulouse 2

Abstract

Tests for continuity of regression function based on local linear estimators are developed and their limit properties are studied. Bootstrap method is proposed to getapproximation for critical values. Simulation study was conducted in order to check how the tests works in finite sample situation.

More precisely, we consider the regression model :

$$Y_i = m(X_i) + \sigma(X_i)\,\varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent identically distributed (i.i.d.) random vectors with $\varepsilon_1, \ldots, \varepsilon_n$ being i.i.d. random errors with

$$E\varepsilon_i = 0, \qquad \operatorname{var}\varepsilon_i = 1, \qquad E|\varepsilon_i|^{2+\delta} < \infty \tag{2}$$

with some $\delta > 1/2$. The regression function $m(\cdot)$, the variance function $\sigma^2(\cdot)$ and the density $f(\cdot)$ of X_i are unknown functions that are supposed to be smooth except possibly a finite number of points. The density $f(\cdot)$ is bounded away from 0 on $(0, 1)$ and is equal to 0 outside of $\langle 0, 1\rangle$. We are interested in the testing problem concerning the smoothness of the regression function m, namely, we want to test

$$H_0 : m \text{ is smooth function on} \langle 0, 1\rangle \tag{3}$$

against

$$H_1 : m \text{ has at least one jump in} (0, 1). \tag{4}$$

The aim of the paper is to develop procedures for testing problem (H_0, H_1) based on nonparametric estimators of $m(\cdot)$ and $\sigma^2(\cdot)$, particularly, based on their locally linear estimators.

Régression pénalisée par ondelettes sous contrainte de monotonie

Anestis ANTONIADIS, Jéremie BIGOT * et Irène GIJBELS

* Adresse pour correspondance :
Laboratoire de Statistique et Probabilités
Université Paul Sabatier, Toulouse
jbigot@cict.fr

Séance du 28 Février 2005

Résumé

Dans cet exposé, nous nous intéressons au problème de la régression nonparamétrique sous contrainte de monotonie. Dans un premier temps, nous proposerons un estimateur basé sur des des techniques de régression pénalisée par ondelettes [1]. Nous montrons que notre approche peut être formalisée comme un problème d'optimisation convexe sous des contraintes linéaires. Des conditions nécessaires et suffisantes de type Karush-Kuhn-Tucker qui garantissent l'existence d'une solution unique sont discutées. On montre que l'estimateur sous contraintes est facilement obtenu via la formulation duale du problème d'optimisation. En particulier, nous montrons que cette approche peut être utilisée pour obtenir un estimateur monotone par régression pénalisée par ondelettes. Nous établissons le taux de convergence de cet estimateur, et nous illustrons ses propriétés sur des échantillons de taille finie à l'aide de simulations. Nous comparons également les performances de notre méthode avec un estimateur sous contrainte de monotonie (basée sur des Splines) qui a été récemment proposé [5].

Le problème de la régression nonparamétrique sous contraintes a été largement étudiée dans le contexte du lissage Spline [2], [3], [4]. Dans une deuxième partie de l'exposé, nous montrons que dans le cas des Splines, ce problème peut se formaliser à l'aide d'outils qui ont été développés pour l'estimation de la déformation entre deux images. Nous proposerons une nouvelle formulation du problème de l'estimation d'une fonction monotone qui ne fait pas intervenir des contraintes basée sur les points du design. Nous nous efforcerons de donner une interprétation

intuitive de ce problème et proposerons quelques pistes de recherche.

Références

[1] Antoniadis, A., Bigot, J. and Gijbels I. (2005). Penalized wavelet monotone regression, preprint.

[2] Gijbels, I. (2004). Monotone regression, to appear in *Encyclopedia of Statistical Sciences, Second Edition*. Editors S. Kotz, N.L. Johnson, C.B. Read, N. Balakrishnan, and B. Vidakovic. Wiley, New York.

[3] Mammen, E. , Marron, J.S. , Turlach, B.A. and Wand, M.P. (2001). A general projection framework for constrained smoothing, *Statist. Sci.*, **16**, 232–248.

[4] Mammen, E. , Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions, *Scand. J. Statist.* , **26**, 239–252.

[5] Zhang, J.-T. (2004). A simple and efficient monotone smoother using smoothing splines, *Journal of Nonparametric Statistics*, **16**, 5, 779–796.

Sur une intégrale tensorielle intrinsèque et ses applications

Alain BOUDOU et Yves ROMAIN

Adresse pour correspondance :
Laboratoire de Statistique et Probabilités, UMR CNRS C55830
118, route de Narbonne, 31062 Toulouse Cedex
boudou@cict.fr romain@cict.fr

Séance du 14 Mars 2005

Résumé

Étant donné deux séries stationnaires indépendantes X_n et Y_n, avec n entier relatif, il est facile de constater que le processus transformé multiplicatif simple $T_n = Y_n X_n$ est aussi stationnaire. Il est bien connu que tout processus stationnaire est transformé de Fourier d'une mesure aléatoire qui le définit d'une façon biunivoque.
Ici, nous nous proposons d'exprimer, en fonction des mesure s aléatoires Z_x et Z_Y respectivement associées à (X_n) et (Y_n), la mesure aléatoire dont la transformée de Fourier est $(X_n Y_n)$. Nous étudions également les divers éléments liés au processus $(X_n Y_n)$ telles que les mesure et densité spectrales.
Pour cela, nous définissons le produit $Z_x \otimes Z_Y$ des mesures aléatoires Z_x et Z_Y et établisssons une formule de type Fubini d'intégration par rapport à cette mesure. On constate alors que la mesure image de $Z_x \otimes Z_Y$ par l'application somme, que nous appellerons naturellement *produit de convolution des mesures aléatoires* Z_x *et* Z_Y est, à une isométrie près, la mesure aléatoire associée au processus $(X_n Y_n)$. Ces résultats, récemment publiés dans une note interne, peu vent concerner par exemple (un des buts de cet exposé étant de trouver d'autres exemples...), les changements d'unité de mesure, les perturbations multiplicatives ou encore des problèmes inverses (recherche d'inconnue) dans l'équation $T_n = Y_n X_n$.

.

Réduction de dimension en régression non paramétrique : vers la dimension infinie

Amadou WADE

Adresse pour correspondance :
Université Paul Sabatier Laboratoire de Statistique et Probabilités
UMR CNRS C55830
118, route de Narbonne, 31062 Toulouse Cedex
et Université Gaston Berger de Saint-Louis Sénégal
wade@cict.fr

Séance du 21 Mars 2005

Résumé

Stone en 1985 a montré que la vitesse optimale de convergence de l'estimateur de la fonction inconnue de régression r dans le cadre de la régression non paramétrique dans $\mathbb{R}^p$, $\quad p \in \mathbb{N}^*$, dépend de la dimension p. Autrement dit quand p augmente la vitesse de convergence devient de plus en plus mauvaise. Ce phénomène est appelé fléau de la dimension. C'est la raison pour laquelle beaucoup de statisticiens se sont investis pour la recherche de solutions à ce fléau de la dimension. Une solution consiste à proposer des modèles de réduction de dimensions. Parmi ces modèles, on peut citer : les modèles additifs, les modèles généralisés-modèles avec interactions, les modèles de projections révélatrices. Pour ce qui concerne le dernier modèle il y'a dautres paramètres supplémentaires qui viennent s'y ajouter notamment les sous espaces réduits $\beta_k, k = 1, .., p$ appelés edr et la dimension q.

En somme des résultats satisfaisants sont obtenus pour ce qui concerne les paramètres à estimer. Notre objectif maintenant est de faire une extrapolation en dimension infinie et nous nous proposons d'étudier le modèle de projection poursuite suivant :

$$Y = \mu + r(\alpha_1 X_1, .. + \alpha_p X_p) + \varepsilon$$

avec $X_i \in (H, d), i = 1, .., p$, H étant un espace de Hilbert, d étant une semi métrique sur H, et où les $\alpha_i, i = 1, .., p$ sont des réels vérifiants :

$$\sum \alpha_i = 1.$$

Références

Andrews,D and Whang, Y(1990) :" Additive interactive regression models : circumvention of the curse of dimensionality", *Econome tric Theory,* 6,466-479.

Buja, A.,Hastie, T et Tibshirani,R. (1989) :"Linear smoothers and additive models," *Ann.Statistics,* 17, 453-455.

Chen,H. (1991) : "estimation of a projection-pursuit Type Regression Model," *The Annals of Statistics,* 19, 142-157.

Hastie, T. and Tibshirani, R. (1987) :" Generalized additive models : some applications," *J. of Amer.Assoc.*,87,371-386.

Horowitz, J.L. (1992) : "A Smoothed Maximum Score Estimator for the Binary Respoonse Model," *Econometrica*, 60, 505-531.

Li,K.C. (1991) "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*,86,316-3 32.

Li,K.C. (1992) : On Principal Hessian Directions for Data Visualisa tion and Dimention Reduction : Another Application of Stein's Lemma," *Journal of the American Statistical Association*,87, 102 5-1040.

Xia,Y, and li,W.K (1999) : "On single Index Coefficient Regression Models," *Journal of the American Statistical Association*, 9 4, 12755-1285.

Asymptotic properties of a nonparametric regression function estimator with randomly truncated data

Elias OULD-SAID

Adresse pour correspondance : Université du Littoral
Laboratoire de Mathématiques Pures et Appliquées
Centre de la mi voix, BP 699
62228 Calais, Cedex.
Elias.Ould-Said@lmpa.univ-littoral.fr

Séance du 4 Avril 2005

Abstract

In this talk, we define a new kernel estimator of the regression function under a left truncation model. We establish the pointwise and uniform strong consistency over a compact set and give a rate of convergence of the estimate. The pointwise asymptotic normality of the estimate is also given. Some simulations are given to show the asymptotic behavior of the estimate in different cases. The distribution function and the covariable's density are also estimated.

.

Régression pour réponses fonctionnelles

Luboš PRCHAL

Adresses pour correspondance :
Département de Statistique, Université Charles à Prague
Sokolovská 83, 186 00 Prague 8, République tchèque
&
Laboratoire de Statistique et Probabilités
Université Paul Sabatier,
118, route de Narbonne, 31062 Toulouse Cedex, France
e-mail : prchal@cict.fr

Séance du 11 Avril 2005

Résumé

On s'intéresse au problémes de l'explication d'une variable fonctionnelle Y par une variable X fonctionnelle ou réelle selon le cas. Dans la première partie, lorsque la variable X est fonctionnelle on considère la régression linéaire fonctionnelle de Y sur X

$$Y(t) = \int_S X(s)\beta(s,t)\,ds + \varepsilon(t),$$

où $\mathsf{E}\varepsilon(t) = 0$ et $\mathsf{E}X(s)\varepsilon(t) = 0$ pour presque tout s et t.

On donne tout d'abord une condition d'existence et d'unicité de β. On introduit ensuite un estimateur de ce paramètre fonctionnel basé sur la méthode des moindres carrés pénalisés utilisant une décomposition de β sur une base de B-splines. La méthode est illustrée par des simulations Monte Carlo.

La seconde partie est consacrée au cas où la variable explicative X est réelle. On considère ainsi la régression m de Y sur X

$$\mathsf{E}[Y(t)|X] = m(t,X),$$

dans un cadre non-paramétrique. On définit l'estimateur a noyau de m étudié par Cardot (2005). On illustre la méthode sur des données de radioactivité atmosphérique. Enfin, on s'intéresse au test de l'hypothèse nulle

$$H_0 : m(t,X) = 0, \quad \forall t.$$

Le niveau et la puissance du test basé sur des permutations est étudié à l'aide de simulations.

Références

Cardot, H. (2005). *Nonparametric regression for functional responses with application to conditional functional principal component analysis.* Preprint.

Sélection de modèles graphiques gaussiens

Hélène MASSAM

Adresse pour correspondance :
Université York, Toronto, Canada.
e-mail : massamh@mathstat.yorku.ca

Séance du 9 Mai 2005

Résumé

Les modèles graphiques gaussiens se sont avérés être très utiles pour l'analyse des données complexes de haute dimension. Ces modèles servent en particulier à l'analyse des données du génome car ils permettent de représenter graphiquement les différentes relations de dépendence entre les gènes.

L'une des difficultés techniques rencontrée lors du processus de sélection de modèles, dans un cadre bayesien, est le calcul d'une constante de normalisation lorsque le graphe du modele est non décomposable. Nous proposonsune solution à ce problème et en déduisons aussi une façon d'échantilloner l'hyper inverse Wishart.

.

Moments et estimation des paramètres pour une loi de Wishart non centrale

Gérard LETAC

Adresse pour correspondance : Université Paul Sabatier
Laboratoire de Statistique et Probabilités, Toulouse
e-mail : letac@cict.fr

Séance du 23 Mai 2005

Résumé

Les lois de Wishart non centrées sont les extensions matricielles des lois de ki-deux non centrées. Elles dépendent pratiquement de trois paramètres : forme, échelle et décentrage. Si le décentrage est connu, on a une famille exponentielle et les techniques d'estimation qui vont avec. La fonction variance se calcule. Les choses deviennent plus intéressantes quand il faut estimer le décentrage, et les travaux sérieux sur ce sujet sont très rares. Comme la densité de ces lois s'exprime comme la somme d'une série de polynômes zonaux, la méthode du maximum de vraisemblance est inaccessible. On procède donc ici par une nouvelle méthode de moments. Auparavant j'expliquerai comment on calcule les moments de la forme

$$\mathbf{E}(\langle h_1, X\rangle \dots \langle h_k, X\rangle)$$

quand $h_1, \dots, h_k$ sont des matrices symétriques ($\langle h_1, X\rangle$ signifie trace de $h_1 X$). Pas d'excitation, pour $k = 4$ c'est déjà une somme de 213 termes. C'est un travail fait avec Hélène Massam.

.

Testing the equality of k regression curves with extension to censored data

Wenceslao GONZALEZ MANTEIGA

Adresse pour correspondance :
Departamento de Estadística e Investigación Operativa
University of Santiago de Compostela
15782, Santiago de Compostela, Espagne
e-mail : wenceslao@usc.es

Séance du 10 Juin 2005

Exceptionnellement couplée avec le
séminaire de l'équipe SMASH de l'Univeristé de Toulouse 2

Abstract

We introduce a new procedure to test the hypothesis of equality of the k regression curves in the context of nonparametric regression. The test is based on the comparison of two estimators of the distribution of the errors in each population. Kolmogorov-Smirnov and Cramer von Mises type statistics are considered, and their asymptotic distributions are obtained. The proposed tests can detect local alternatives converging to the null hypothesis at the rate $n^{-1/2}$. We describe a bootstrap procedure in order to approximate the critical values and present the results of a simulation study. The method is extended to regression models with censored responses.

.

PBLICATIONS ANTÉRIEURES

Tous les documents ci-dessous sont disponibles en ligne à l'adresse
http ://www.lsp.ups-tlse.fr/
dans la rubrique "rapports Techniques"

- 3èmes Journées de Statistique Fonctionnelle et Opératorielle : Toulouse 13-14 Juin 2005. Publications du # **LSP 2005-05**.

- Résumés des activités 2003-2004. Publications du # **LSP 2004-06**.

- Résumés des activités 2002-2003. Publications du # **LSP 2003-05**.

- 2èmes Journées de Statistique Fonctionnelle : Toulouse 12-13 Juin 2003. Publications du # **LSP 2003-06**.

- Résumés des activités 2001-2002. Publications du # **LSP 2002-12**.

- Journées de Statistique Fonctionnelle : Toulouse 10-11 Juin 2002. Publications du # **LSP 2002-09**.

- Résumés des activités 2000-2001. Publications du # **LSP 2001-07**.

- Résumés des activités 1999-2000. Publications du # **LSP 2001-05**.