Groupe de travail STAPH: Statistique Fonctionnelle

Partie II: Recueil de résumés 2000-2001

Coordinateurs
H. Cardot, F. Ferraty, Y. Romain, P. Sarda et P. Vieu

TABLE DES MATIERES

Rappel des exposés précédents

Année 1999-2000. Sommaire de la publication du Laboratoire de Statistique et Probabilités de Toulouse **LSP 2001-XX**

Estimation fonctionnelle, par P. Sarda et P. Vieu
Modélisation pour variables fonctionnelles dans un contexte explicatif, par H. Cardot et F. Ferraty
Sur et pour une approche fonctionnelle en statistique, par Y. Romain
Produit de convolution de mesures spectrales, par A. Boudou
The geometrical theory of estimating functions, par C. Small
Inférence statistique pour des estimateurs de discontinuités dans un cadre non paramétrique, par V. Couallier
Nonparametric estimation in null recurrent time series par D. Tjøstheim
ACP de fonctions de densité. Application aux données climatiques, par T. Antoniadou *et al.*
Modèle non linéaire fonctionnel: une approche par régression inverse, par A.F. Yao et L. Ferré
Estimation bayésienne de l'intensité d'un processus de Cox non homogène par une méthode MCMC à saut réversible, par M. Goulard
Permutation tests in change point analysis, par J. Antoch et M. Hušková
Inférence statistique pour la localisation d'une discontinuité par régression linéaire locale, par G. Grégoire
Non causalité et discrétisation fonctionnelle, théorèmes limites pour un processus ARHX(1), par S. Guillas
Data exploration using piecewise polynomial regression trees, par P. Chaudhuri

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. Rappelons que ce groupe de travail en Statistique Fonctionnelle, créé il y a deux ans 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 l'ouverture de la démarche par la grande diversité des exposants. En préambule de ce document, un court texte est présenté afin de tirer le bilan des deux premières années de travail et afin surtout de mieux préparer l'avenir en faisant perdurer cette dynamique de recherche.

Abstract We present the abstracts (of size more or less important acording to the wishes of their authors) of the several different talks given during the sessions of the working group STAPH. This group in Functional Statistics is born two 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 foreword, a short text is presented to take the stock of the activities of this group during its two first years of existency in order to make an efficient preparation of the future.

Sumario This 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. Este grupo de trabajo en el campo de Estadistica Funcional ha sido creado hace dos 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 principio del documento, empezamos con un corto texto de presentacion en lo cual hacemos un chequeo de las actividades de este grupo desde dos anos y en lo cual planteamos los fundamentos para el proximo futuro.

STAPH: Bilan et perspectives aprés deux années d'existence

Hervé Cardot, Frédéric Ferraty
Yves Romain, Pascal Sarda et Philippe Vieu

Co-fondateurs et coordinateurs du groupe de travail STAPH
Laboratoire de Statistique et Probabilités

cardot@toulouse.inra.fr, ferraty@cict.fr, romain@cict.fr
sarda@cict.fr, vieu@cict.fr

Aprés deux années d'existence au sein du **Laboratoire de Statistique et Probabilités de Toulouse**, il nous a semblé souhaitable de tirer un bref bilan de nos activités afin de mieux maintenir et d'actualiser en permanence la dynamique existante autour de la **Statistique Fonctionnelle**. Ce bilan est en particulier le fruit de nombreuses discussions avec divers participants à ce groupe de travail.

Avec une fréquence moyenne de deux séances par mois, et une participation moyenne d'une quinzaine de personnes (participation sans cesse renouvelée), nous pensons avoir atteint le premier de nos objectifs, à savoir la création d'un lien scientifique permanent autour de la Statistique Fonctionnelle à Toulouse. Ce succés est le témoin, au même titre que les Habilitations récentes de nos collègues Yves Romain et Pascal Sarda, au même titre que les doctorats récents de Vincent Couallier et Christine Camlong, et au même titre que les deux cours enseignés en 2000-2001 et 2001-2002 au DEA de Mathématiques Appliquées de Toulouse, d'une dynamique locale incontournable autour des divers aspects fonctionnels de la Statistique moderne. Non seulement cet engouement dépasse largement le cadre toulousain mais il s'inscrit à part entière dans une dynamique qui secoue de manière générale la scène internationale des statisticiens comme en témoigne la grande diversité des 32 séances de travail que nous avons vécues depuis deux ans, qu'il s'agisse de diversité dans l'origine géographique de nos intervenants (environ 1/3 de toulousains, 1/3 d'extérieurs français et 1/3 d'intervenants étrangers) ou bien de diversité dans la nature de leurs interventions (exposés plutôt didactiques, présentation de nouveautés théoriques ou/et de recherches appliquées). Que tous ces intervenants soient ici remerciés, et plus généralement tous ceux qui ont participé à nos séances de travail ou tout simplement soutenu notre démarche.

D'un point de vue scientifique, il est en particulier un point important que nous souhaitons mettre en évidence afin de mieux prépaper l'avenir. Il s'agit du fait que les activités de notre groupe de travail ont permis la réalisation de plusieurs travaux autour des deux thèmes fondateurs de notre démarche (et de

leurs intersections) à savoir la **Statistique Opératorielle** et la **Statistique Nonparamétrique** (voir en particulier les exposés (1), (2) et (17)). Dans un futur immédiat nous souhaitons poursuivre autour de ces deux thèmes, et nous souhaitons encourager tous les statisticiens (toulousains ou non) intéressés par cela à nous contacter afin de pouvoir voir comment les associer concrètement à une ou plusieurs séances à venir. Toujours dans cet esprit là, il nous semble que la lisibilité de notre démarche n'a pas été toujours perçue comme nous le souhaitions, en ce sens que l'appellation Statistique Fonctionnelle peut souvent apparaître comme une autre appelation de la Statistique Nonparamétrique ou de l'Estimation Fonctionnelle. Il nous semble nécessaire de réaffirmer avec force qu'il n'en est rien et que l'Estimation Fonctionnelle n'est qu'une facette de notre démarche. Nous sommes fortement convaincus que l'Estimation Fonctionnelle elle-même ne pourra se développer qu'en gardant et multipliant les passerelles avec les autres aspects fonctionnels de la Statistique et en particulier avec la Statistique Opératorielle. Nous sommes aussi convaincus que, outre ses apports immédiats à la Statistique Nonparamétrique, la Statistique Opératorielle est une porte entrouverte vers de nombreux autres domaines scientifiques (dont, par exemple, la Statistique quantique et les Mathématiques non commutatives, ou encore les Mesures tensorielles aléatoires et spectrales associées à des processus stationnaires...). En conséquence, et pour assurer de manière plus claire la lisibilité de notre démarche et assurer ainsi son avenir, nous avons décidé de modifier l'intitulé de notre groupe en l'appelant désormais **STAPH** groupe de travail en **Statistique Fonctionnelle et Opératorielle**.

Pour ce qui concerne le déroulement futur de nos séances nous pensons continuer dans l'optique de deux séances par mois, sans aller au delà vu le nombre déjà élevé de séances de séminaires en Statistique disponibles sur la place toulousaine. Nous souhaitons aussi essayer de nous éloigner de la forme académique de type séminaire au profit se séances plus axées sur des échanges entre l'orateur et le public. Pour terminer, remercions encore tous les orateurs et tous les participants à nos séances de travail, et souhaitons de bonnes vacances à toutes et à tous ...

Toulouse, le 30 Juin 2001.

Sur les effets de la dimension en estimation fonctionnelle: du réel vers le fonctionnel

Philippe VIEU

Laboratoire de Statistique et Probabilités
Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse Cedex, France
vieu@cict.fr

Exposé du 16 Octobre 2000

---

1. Introduction

L'objectif principal (presque unique) de l'exposé est de discuter des effets de la dimension en estimation fonctionnelle. La discussion sera menée à partir de l'étude des vitesses de convergence d'estimateurs de la régression dans des modèles non-paramétriques

$$Y = r(X) + \epsilon.$$

Dans tous les modèles que nous étudierons, Y et ϵ seront des variables aléatoires réelles, et nous distinguerons trois cas de figure selon que:

- X est réelle;
- X est vectorielle;
- X est à valeurs dans un espace de dimension infinie.

Les modèles non-paramétriques se caractérisent par une hypothèse de régularité sur l'objet à estimer r. Nous limiterons notre propos à une hypothèse de continuité sur sur r ou tout au plus à une condition de régularité de type Lipschitz

$$|r(u) - r(v)| \leq C||u - v||^{\beta}, \ (u, v) \in E^2, \tag{1}$$

E étant l'espace sur lequel r prend ses valeurs dont on supposera au minimum qu'il s'agit d'un espace vectoriel semi-normé, et $||.||$ désignant cette semi-norme sur E.

Pour simplifier l'exposé, dans tous les cas nous nous limiterons à un cadre d'estimation ponctuelle, c'est à dire à l'estimation de $r(x)$, x étant un point fixé de E.

Concernant la forme de l'exposé, lors des paragraphes 2 et 3 qui concernent les cadres réels et vectoriels, il s'agira plutôt de rappeler quelques résultats bien connus, tandis que lors du paragraphe 4 consacré au cadre fonctionnel nous présenterons un résultat récent issu de Ferraty et Vieu (2000).

2. Le cadre réel

Dans un premier temps nous regardons le cas où X est réelle. Nous rappèlerons rapidement des résultats bien connus en ce domaine, à savoir que les vitesses optimales de convergence presque sûre sous un modèle (1), et pourvu que X admette une densité f par rapport à la mesure de Lebesgues sur $\mathbb{R}^p$ qui soit elle aussi Lipschitzienne, sont en

$$\left(\frac{n}{logn}\right)^{-\frac{\beta}{2\beta+1}}. \tag{2}$$

Nous montrerons rapidement que les estimateurs à noyau de convolution atteignent, par un bon choix du paramètre de lissage, ces vitesses optimales. Les résultats que nous rappelerons sont bien connus. La plupart sont dûs à Collomb (1976) et ils ont été abondament repris et/ou améliorés depuis. Nous renvoyons à Sarda et Vieu (2000) ou à l'exposé de Sarda (2000) pour de plus amples références sur les propriétés de ces estimateurs de convolution, et à Stone (1982) pour ce qui concerne l'optimalité des vitesses (2).

3. Le cadre vectoriel

Dans un deuxième temps nous regarderons le cas où X est à valeurs dans $\mathbb{R}^p$. Nous montrerons, là aussi trés rapidement, comment les résultats précédents s'étendent sans difficultés du cadre réel au cadre vectoriel. Nous verrons que les estimateurs à noyau de convolution atteignent des vitesses de convergence presque sûre sous un modèle (1) qui sont en

$$\left(\frac{n}{logn}\right)^{-\frac{\beta}{2\beta+p}}, \tag{3}$$

vitesses dont Stone (1982) a aussi établi l'optimalité sous l'hypothèse (1) pourvu que X admette une densité f par rapport à la mesure de Lebesgues sur $\mathbb{R}^p$ qui soit elle aussi Lipschitzienne.

Là aussi il s'agit de résultats bien connus, dont l'origine remonte à Collomb (1976), et nous nous bornerons à discuter l'influence de la dimension p, en insistant sur le fait qu'une solution à la relative lenteur des vitesses (3) est à chercher dans la construction d'autres modèles (puis d'autres estimateurs sous ce modèle bien sûr), et non pas dans la construction d'autres estimateurs pour le modèle (1). Nous décrirons quelques uns de ces modèles que nous appelons "modèles non-paramétriques pour réduction de dimension". Nous renvoyons à Stone (1985), Pelegrina *et al.* (1996) ou Schimek (2000) pour plus de références sur ce type de modèles, et à Camlong (2000) pour une étude détaillée d'estimateurs de type noyau sous ces modèles.

4. Le cadre fonctionnel

Pour terminer nous regarderons le cas où X est à valeurs dans un espace vectoriel semi-normé quelconque $(E, ||.||)$. Il s'agit d'un problème pour lequel les connaissances sont nettement avancées que dans les deux problèmes précédents. On définira des estimateurs de type convolution adaptés à ce type de problème fonctionnel, et on exposera un résultat récent de Ferray et Vieu (2000).

On supposera que la variable X vérifie, pour x fixé dans E, la condition suivante:

$$lim_{\alpha \to 0^+} \frac{P(||X - x|| \leq \alpha)}{\alpha^a} = c(x), \tag{4}$$

où a et $c(x)$ sont deux réels strictement positifs, ou bien on supposera la condition plus forte suivante:

$$P(||X - x|| \leq \alpha) = \alpha^a c(x) + O(\alpha^{a+\beta}). \tag{5}$$

Nous montrerons alors que sous (1) et (4) on a la convergence presque sûre de nos estimateurs, tandis que sous la condition additionnelle (5) on a des vitesses de convergence en

$$\left(\frac{n}{logn}\right)^{-\frac{\beta}{2\beta+a}}. \tag{6}$$

Ce résultat sera ensuite discuté et on montrera en particulier comment il permet de retrouver ceux énoncés dans les paragraphes précédents lorsque X était réelle ou vectorielle. On discutera aussi les problèmes que laisse ouverts ce résultat, et notament ceux liés à la comparaison entre les hypothèses fractales du type de celles que nous utilisons ici, c'est à dire du type (4) ou (5), et des hypothèses comme celles utilisées par exemple par Pesin (1993) ou Bardet (1997) et qui s'écrivent plutôt sous la forme

$$lim_{\alpha \to 0^+} \frac{logP(||X - x|| \leq \alpha)}{log\alpha} = a. \tag{7}$$

5. Une application

Pour terminer, on présentera un exemple de jeu données spectrométrique pour lequel la modèlisation et les estimateurs du paragraphe 4 sont intéressants. Cet exemple nous servira à mettre en évidence l'intérêt de travailler avec des espaces vectoriels semi-normés plutôt qu'avec des espaces Hilbertiens. Il nous permettra aussi de voir que, comme cela est bien connu dans les cas réels et vectoriels, les estimateurs non-paramétriques (et en particulier ceux de type noyau), sont intéressants en tant qu'approche exploratoire.

Références

Bardet, J.M. (1997). Tests d'autosimilarité des processus gaussiens. Dimension fractale et dimension de corrélation. *Thèse 3eme cycle, Paris-Sud.*

Camlong, C. (2000). Intégration marginale pour modèles additifs de régression multidimensionnelle. *Thèse, Toulouse 3, à soutenir en 2000.*

Collomb, G. (1976). Estimation non-paramétrique de la régression par la méthode du noyau. *Thèse 3eme cycle, Toulouse 3.*

Pelegrina, L., Sarda, P. et Vieu, P. (1996). On multidimensional nonparametric regression. *In Proceedings of Computational Statistics, Ed. A. Prat*, Physica-Verlag.

Pesin, Y.B. (1993). On rigourous mathematical definitions of correlation dimension and generalized spectrum for dimensions. *J. statist. Phys.*, **71**, 529-547.

Sarda, P. (2000). Estimation non-paramétrique. *Groupe de travail STAPH 1999-2000, Pub. du Labo. Statist. Prob., Toulouse 3.*

Sarda, P. et Vieu, P. (2000). Kernel Regression. *Smoothing and Regression: Approaches, computation, and application, Ed. M.G. Schimek*, 43-70, Wiley Series in Probability and Statistics.

Schimek, M. (2000). *Smoothing and Regression: Approaches, computation, and application, Ed. M.G. Schimek*, Wiley Series in Probability and Statistics.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

Stone, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **85**, 689-705.

Estimations dans le modèle linéaire fonctionnel

Frédéric FERRATY
En collaboration avec Hervé CARDOT et Pascal SARDA

Laboratoire de Statistique et Probabilités
Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse Cedex, France
ferraty@cict.fr, cardot@toulouse.inra.fr, sarda@cict.fr

Exposé du 23 Octobre 2000

---

1. Introduction et présentation du modèle

L'objectif de cet exposé est de présenter un modèle de régression linéaire pour lequel la variable réponse Y est une variable aléatoire réelle alors que le régresseur est une variable aléatoire X définie sur $(\Omega, \mathcal{A}, P)$ et à valeurs dans un espace fonctionnel H de dimension **infinie**. On parle alors de variable aléatoire fonctionnelle (v.a.f.) ainsi que de modèle de régression linéaire fonctionnel. Par ailleurs, notons que ce modèle est "doublement" fonctionnel en ce sens que l'objet (opérateur linéaire continu de H dans $\mathbb{R}$ ou élément de H) que l'on souhaite estimer est aussi de nature fonctionnelle. Du point de vue de l'individu statistique, il peut être :

- soit une courbe (de croissance, de température, ...),
- soit une fonction de plusieurs variables (champ de pression, ...),...

qui est donc la réalisation d'une v.a.f. X à valeurs dans H. A titre d'exemple, on peut particulariser H :

- H = espace des fonctions de carré intégrable,
- H = espace d'opérateurs, ...

On observe ainsi

$$(X_e)_{e\in E} = \{X(e) \; ; \; e \in E\}$$

où $E = I \subset \mathbb{R}, \mathbb{C}, \ldots$. Notons que le cas particulier $(X_t)_{t\in T}$ où t désigne le temps $(T \subset \mathbb{R})$ correspond à l'étude d'un processus à temps continu. Il va sans dire que les applications sont extrêment nombreuses ; pour s'en convaincre, il suffit de donner les quelques exemples issus de différents domaines :

- **agronomie** : quelle influence peut avoir la pluviométrie et les courbes de température sur le rendement d'une production céréalière?

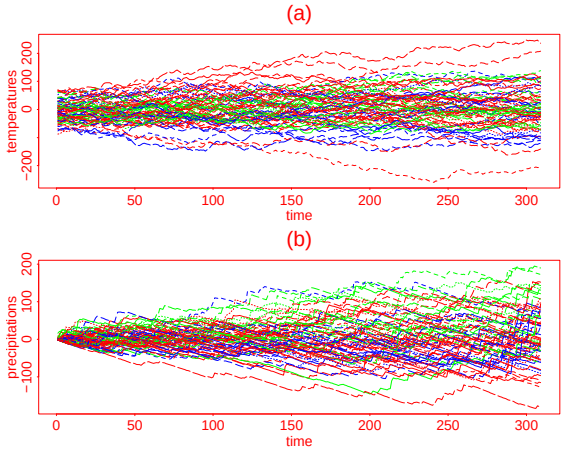

Figure 1: Courbes de pluviométrie et de température ; relevés journaliers cumulés centrés.

- **chimie quantitative** : comment estimer le taux de lipide contenu dans un élément à partir de son spectre dans le proche infra-rouge?

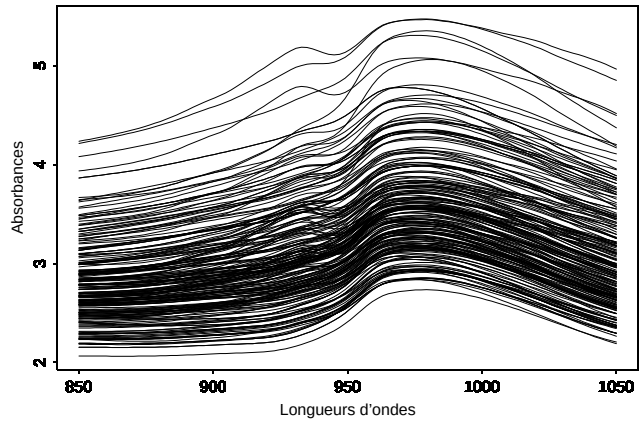

Figure 2: Courbes spectrométriques.

- **économique** : comment estimer les valeurs futures d'un indice boursier à partir de son évolution passée?

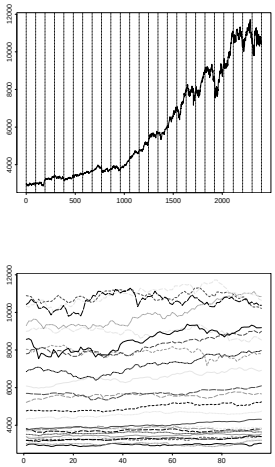

Figure 3: Le découpage en sous-intervalles nous ramène à l'observations de plusieurs courbes.

Dans ce qui suit, on se restreint au cas où $H = L^2_{[0,1]}$ est un espace hilbertien séparable muni du produit scalaire (resp. de la norme) $< .,. >$ (resp. $\|.\|$). On associe à H la tribu des boréliens $\mathcal{B}_H$ et on note

$$L^2_H(P) = \left\{ Z : (\Omega, A, P) \to (H, \mathcal{B}_H) \ / \ \mathbb{E}\left(\|Z\|^2\right) < +\infty \right\}.$$

L'utilisation de telles v.a.f. nécessite certaines définitions élémentaires. En particulier, l'**espérance** de X est l'élément μ de H défini par :

$$< \mu, x > = \ \mathbb{E}(< X, x >), \qquad x \in H$$

et μ est noté $\mathbb{E}X$. De plus, sachant que $x \otimes y(u) \ = < u, x > y$, $\forall (x, u) \in H \times H$, on appelle **opérateur de covariance de** X, l'opérateur noté Γ de H dans H défini par :

$$\Gamma x \ = \ \mathbb{E}((X - \mu) \otimes (X - \mu)(x)),$$

et si X est centré ($\mu = 0$) :

$$\Gamma x \ = \ \mathbb{E}(X \otimes X(x)) \ = \ \mathbb{E}(< x, X > X).$$

Lorsque $X \in L^2_H(P)$, l'opérateur Γ possède de "bonnes" propriétés et en particulier, Γ est un opérateur symétrique, positif et nucléaire (pour plus de précisions, voir Dauxois et Pousse (1976), Romain (1979), Fine (1981), Dauxois-Pousse-Romain (1982)). Ajoutons que l'on dispose d'outils performant concernant la manipulation des opérateurs linéaires (Nagy et Riesz (1952), Dunford et Schwartz (1963), Gohberg et Krejn (1971), Kato (1976), Chatelin (1983), ...). On suppose donc que $X \in L^2_H(P)$ avec $H = L^2_{[0,1]}$ Hilbert séparable et Y une v.a **réelle** définie

sur $(\Omega, \mathcal{A}, P)$. On dispose ainsi d'un échantillon de taille n, $(X_i, Y_i)_{i=1,\ldots,n}$, n variables i.i.d. de même loi que (X, Y) et le modèle de régression linéaire fonctionnel (*Cf.* Hastie et Mallows (1993), Ramsay et Silverman (1997), Cardot, Ferraty et Sarda (1999-2000)) est défini par

$$Y_i \;=\; \int_0^1 \psi(t) X_i(t) dt \;+\; \varepsilon_i, \quad i = 1, \ldots, n$$

où $\psi \in H$, $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2$ et ε_i indépendant de X_i. De manière équivalente, on peut écrire aussi :

$$Y_i \;=\; \Psi\left(X_i\right) \;+\; \varepsilon_i, \quad i = 1, \ldots, n$$

où Ψ est un opérateur linéaire continu de H dans $\mathbb{R}$.

Objectif : estimer ψ ou Ψ.

2. Estimation empirique de Ψ et propriétés asymptotiques

L'estimation de l'opérateur Ψ est alors basée sur la relation suivante

$$\boldsymbol{\Delta} \;=\; \Psi \boldsymbol{\Gamma},$$

où $\boldsymbol{\Delta} \;=\; \mathbb{E}\left(X \otimes Y\right)$. Soit $\boldsymbol{\Gamma}_n \;=\; \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$ (resp. $\boldsymbol{\Delta}_n \;=\; \frac{1}{n}\sum_{i=1}^n X_i \otimes Y_i$) l'opérateur de covariance (resp. covariance croisée) empirique. Considérons H_K le sous-espace vectoriel de H engendré par les K fonctions propres associées aux K plus grandes valeurs propres de Γ_n et Π_K la projection orthogonale sur H_K. L'estimateur de Ψ est obtenu en projetant les données sur H_K et en inversant Γ_n dans cet espace :

$$\Psi_n \;=\; \Delta_n \Pi_K (\Pi_K \Gamma_n \Pi_n)^{-1}.$$

On réalise donc une régression linéaire multiple de Y sur les K variables explicatives :

$$< X, V_{1,n} >, \ldots, < X, V_{K,n} >,$$

où $(V_{j,n})_{j=1}^K$ sont les fonctions propres associées aux plus grandes valeurs propres $(\lambda_{j,n})_{j=1}^K$ de Γ_n. On dit alors qu'on réalise la régression sur les K premières composantes principales fonctionnelles.

Théorème 1. Convergence en probabilité de Ψ_n
Hypothèses

(H.1) Les valeurs propres de Γ *sont distinctes et non nulles;*

(H.2) $\mathbb{E}\|X\|^4 \;<\; +\infty$*;*

(H.3) la suite $K = K_n$ tend vers l'infini et vérifie :

$$\begin{cases} \lim\limits_{n\to+\infty} n\lambda_{K_n}^4 = +\infty; \\ \lim\limits_{n\to+\infty} \dfrac{n\lambda_{K_n}^2}{\left(\sum_{j=1}^{K_n} a_j\right)^2} = +\infty, \end{cases}$$

où

$$a_j = \begin{cases} \dfrac{2\sqrt{2}}{\lambda_1 - \lambda_2}, & si\ j = 1; \\ \dfrac{2\sqrt{2}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} & sinon \end{cases}$$

$$\sup_{\|x\|=1} |\Psi_n(x) - \Psi(x)| \longrightarrow 0, \ \textit{en proba, quand } n \to +\infty.$$

Remarque. Il suffit de remplacer $(H.3)$ par :

$$\text{(H'.3)} \quad \begin{cases} \lim\limits_{n\to+\infty} n\lambda_{K_n}^4 / log\, n = +\infty; \\ \lim\limits_{n\to+\infty} \dfrac{n\lambda_{K_n}^2}{\left(\sum_{j=1}^{K_n} a_j\right)^2 log\, n} = +\infty, \end{cases}$$

et de supposer que X et ε sont presque sûrement bornées pour avoir la convergence presque sûre.

Par ailleurs, il apparaît, suivant les situations, que l'estimateur empirique que nous proposons manque de régularité lorsqu'on considère des échantillons de "petite taille". Ceci nous conduit à proposer un estimateur plus lisse.

3. Régression sur composantes principales lisses

Dans le but d'obtenir une version lisse de l'estimateur empirique de Ψ, on introduit les fonctions Splines (De Boor, 1978, Schumaker, 1981). On considère sur $[0,1]$, une subdivision en k sous-intervalles $t_1 = 0, t_2, \ldots, t_{k+1} = 1$ (noeuds) et soit q un entier positif. Soit $S_{q,k}$, l'ensemble des fonctions s telles que :

- s est un polynôme de degré q sur $[t_j, t_{j+1}]$, $j = 1, \ldots, k$;
- s est $q-1$ fois continûment dérivable sur $[0,1]$.

L'ensemble $S_{q,k}$ est de dimension $q+k$ et les fonctions B-splines $\{B_l\}_{l=1,\ldots,k+q}$ en constituent une base.
La régression sur composantes principales lisse procède de la manière suivante :
Étape 1 : lissage des courbes ; pour $i = 1, \ldots, n$, on pose :

$$\widehat{X}_i = \arg\min_{\tilde{X}\in S_{q,k}} \left\{ \int_0^1 \left(\tilde{X}(t) - X_i(t)\right)^2 dt \right\}.$$

Étape 2 : On réalise la régression sur composantes principales fonctionnelles comme précédemment en remplaçant X_i par $\widehat{X}_i$:

$$\widehat{\Psi}_n \;=\; \widehat{\Delta}_n \widehat{\Pi}_K \left(\widehat{\Pi}_K \widehat{\Gamma}_n \widehat{\Pi}_n \right)^{-1}$$

où $\widehat{\Gamma}_n \;=\; \frac{1}{n} \sum_{i=1}^n \widehat{X}_i \otimes \widehat{X}_i$ et $\widehat{\Delta}_n \;=\; \frac{1}{n} \sum_{i=1}^n \widehat{X}_i \otimes Y_i$.

Théorème 2. Convergence en probabilité de $\widehat{\Psi}_n$
Hypothèses
(H.1) et (H.3) sont vétifiées,
(H.4) $X \in C^l([0,1])$ *p.s. et* $\mathbb{E} \left\| X^{(l)} \right\|^2 < \infty$

$$\|\widehat{\Psi}_n - \Psi\|_2^2 = \mathbb{E} \left(\widehat{\Psi}_n(X) - \Psi(X) \right)^2 \;\to\; 0, \; \textit{en proba., quand } n \to +\infty.$$

4. Estimation du coefficient fonctionnel ψ

Soit $\mathbf{B}(t) = {}^t\,(B_1(t), \ldots, B_{k+q}(t))$ le vecteur des B-splines évaluées au point t. On estime alors ψ en un point t de $[0,1]$ par :

$$\psi_{PBS}(t) \;= {}^t\,\boldsymbol{\psi}_{PBS} \mathbf{B}(t),$$

où ${}^t\boldsymbol{\psi}_{PBS}$ est le vecteur de $\mathbb{R}^{k+q}$ minimisant :

$$\sum_{i=1}^n (Y_i - < {}^t\,\boldsymbol{\psi}\mathbf{B}, X_i >)^2 \;+\; \lambda \int_0^1 \left[({}^t\boldsymbol{\psi}\mathbf{B}(u))'' \right]^2 du.$$

Dans cette situation, nous obtenons à nouveau des résultats asymptotiques.

5. Conclusion

Nous avons vu comment il était possible de traiter un problème de régression linéaire lorsque le régresseur est une v.a.f. Du point de vue théorique, les outils utilisés concerne les opérateurs linéaires, les fonctions splines ainsi que les outils probabilistes inhérents aux v.a.f. D'un point de vue pratique, les deux principales méthodes présentées sont complémentaires. En effet, la régression sur composantes principales lisses (et donc l'estimation de l'opérateur Ψ) suppose la régularité des trajectoires alors que l'estimation du coefficient fonctionnel ψ nécessite une certaine régularité de ce dernier. Ces deux procédures ne sont donc pas en compétition ; suivant les cas de figure, l'une s'avèrera meilleure que l'autre et réciproquement.

Références

Bosq, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes. In Roussas, G., editor, *Nonparametric Functional Estimation and Related Topics*, NATO, ASI Series, 509-529.
Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional Linear Model. *Statist. & Prob. Letters*, **45**, 11-22.
Cardot, H., Ferraty, F. et Sarda, P. (2000). Étude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. *C. R. Acad. Sci. Paris*, t. 330, Série I, 501-504.
Chatelin, F. (1983). *Spectral Approximation of Linear Operators.* Academic, New-York.

Dauxois, J. et Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Thèse de l'Université Paul Sabatier, Toulouse.
Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function: some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136-154.
de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New-York.

Dunford, N. et Schwartz, J. (1963). Linear operators. Interscience Publishers, New-York.
Fine, J. (1981). Analyses en composantes réduites à l'aide de la théorie des perturbations et de représentations matricielles. Thèse de l'Université Paul Sabatier, Toulouse.
Gohberg, I.C. et Krejn, M.G. (1971). Introduction à la théorie des opérateurs linéaires non auto-adjoints dans un espace hilbertien. Dunod, Paris, 1971.
Hastie, T. and Mallows, C. (1993). A discussion of "A Statistical View of Some Chemometrics Regression Tools" by I.E. Frank and J.H. Friedman. *Technometrics*, **35**, 140-143.
Kato, T. (1976). Perturbation theory for linear operators, Springer.
Nagy, B. et Riesz, F. (1965). Leçons d'analyse fonctionnelle. Gauthier-Villars, Paris.
Ramsay, J. O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag.
Romain, Y. (1979). Etude asymptotique des approximations par échantillonnage de l'analyse en composantes principales d'une fonction aléatoire. Quelques applications. Thèse de l'Université Paul Sabatier, Toulouse.
Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley-Interscience.

Differential equations and inverse problems

Anne VANHEMS

Gremaq et Crest, Université Toulouse 1
Manufacture des Tabacs, 21, allées de Brienne
31 000 Toulouse, France
anne.vanhems@univ-tlse1.fr

Exposé du 13 Novembre 2000

Abstract

Let us consider independent identically distributed observations which admit an unknown cumulative density function F. Structural econometrics considers implicit transformation of F. Of course, the theory developped depends on the nature and properties of the transformation considered. For example, there exists a wide literature about integral transformations, like additive models or instrumental variables theory.

We are interested in studying differential transformations of F, and by extension of the conditional expectation. Such a purpose can be justified first by the numerous applications in economics. For example, in a microeconomic context context, Hausman and Newey (1995) estimate nonparametrically the exact consumer surplus by solving a differential equation. In physics, Florens and Vanhems (2000) find the estimator of the ionosphere thickness by solving a differential equation given by physics theory. In finance, Aït-Sahalia (1996) studied the prices of derivative securities by solving a Cauchy problem.

Second, the properties of solutions of such systems are atttractive. We all know that integrating a nonparametric estimator will improve its properties. More generally, we want to examine the properties of integral operators on a nonparametric estimator, and to show that it will improve it.

We intend to replace our study in the context of inverse problems as treated by Tikhonov and Arsenin (1977) and to link it with economic issues.

Références

Aït-Sahalia, Y. (1993). The delta and bootstrap methods for non parametric kernel functionals. *Preprint.*

Ait-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica*, **64**, 527-560.

Bosq, D. (1996). Non parametric statistics for stochastic processus. *Springer-Verlag.*

Choquet, G. (1964). Cours d'analyse tome II Topologie. *Masson et C^{ie}.*

Florens, J.P. and Vanhems, A. (2000). Nonparametric estimation of ionosphere thickness. *Preprint.*

Hausman, J. and Newey, W. K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, **63**, 1445-1476.

Kunze, H. E. and Vrscay, E. R. (1999). Solving inverse problems for ordinary differential equations using the Picard contraction mapping. *Inverse Problems*, **15**, 745-770.

Sibony, M. and Mardon, J.-Cl. (1984). Approximations et equations differentielles. *Hermann.*

Tikhonov, A. N. and Arsenin, V. Y. (1977). Solutions of ill-posed problems. *V.H. Winston & Sons.*

Quelques aspects des grandes déviations en estimation fonctionnelle

Djamal LOUANI

Université de Paris II & Université de Paris V
louani@ccr.jussieu.fr

Exposé du 27 Novembre 2000

Résumé

Nous établissons des résultats de grandes déviations de type Chernoff en estimation fonctionnelle non paramétrique et nous déduisons quelques applications de ceux-ci. Dans le cadre de suites de variables aléatoires réelles indépendantes et identiquement distribuées, nous étudions les déviations ponctuelle, uniforme et en norme L_1 par rapport à la densité sous-jacente de l'estimateur à noyau de la densité de probabilité. Les fonctions de taux sont complétement identifiées. Dans le cas uniforme, à quelques conditions de régularité complémentaires près portant sur la fenêtre de lissage et les queues de la distribution en question, le résultat obtenu est identique à celui établi dans le cas ponctuel pris au point mode de la densité considérée. Dans le cas de la norme L_1, nos résultats sont universels dans le sens où les fonctions de taux sont indépendantes de la densité sous-jacente et du noyau utilisé. Nous obtenons des résultats similaires pour les déviations ponctuelle et uniforme de l'estimateur de la densité par la méthode des séries orthogonales. Nous proposons des applications de nos résultats à l'étude de l'efficacité, au sens de Bahadur, des tests d'hypothèses où diverses comparaisons sont faites et à la selection de modèles illustrée ici par la comparaison des performances de la méthode du noyau et de la méthode des séries orthogonale en estimation de la densité.

Références

Bahadur R.R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia, Pennsylvania.
Dembo A. & Zeitouni O. (1998). *Large deviations techniques and applications.*

Second edition, Springer-Verlag, New York.
Louani D. (1998). Large deviations limit theorems for the kernel density estimator. *Scand. J. Statist.* **25**, *243-253.*
Louani D. (2000). Large deviations for the L_1-distance in kernel density estimation. *J. Statist. Plan. Inf.* **90**, *177-182.*
Louani D. (2001). Large deviations theorems for orthogonal series density estimators and some applications. *Mathematical Methods of Statistics*, À paraître.
Nikitin Ya. (1995). *Asymptotic efficiency of nonparametric tests.* Cambridge University Press.

Non-Uniformity of Job-Matching in a Transition Economy. A Nonparametric Analysis for the Czech Republic.

Stefan SPERLICH *
En collaboration avec Stefan PROFIT

* Adresse pour correspondance:
Departamento de Estadistica y Econometria, Université Carlos III, Madrid
stefan@est-econ.uc3m.es

Exposé du 27 Novembre 2000

1. Abstract

The emergence of open unemployment in central and eastern Europe during the transformation process has created the need to establish modern institutions which provide a framework for worker and job flows on these newly created marketplaces. The question, whether job and worker reallocation in transition economies has evolved to exhibit a similar pattern known from western European labor markets, has been subject to extensive research in recent years. Empirical investigations of the aggregate matching have frequently been used in this context (Boeri and Burda (1996), Burda and Profit (1996), Münch *et al.* (1995)).

Most previous studies have failed to account sufficiently for the heterogeneity of matching technologies: differences may not only appear in regional and district fixed effects but also in marginal effects of the matching factors. In addition, it is plausible that labor market reforms in transition economies have not evolved uniformly since the outset of the transformation period, and returns to scale may vary geographically and over time. Considering this heterogeneity in the matching technology is important, since a misspecified model renders misleading empirical results. Such flexibility enables us to evaluate the local properties of the job-matching. For example, finding locally increasing returns to scale for certain regions and periods, even with constant and decreasing returns to scale on aggregate, may induce multiple equilibria. The main contribution of this study is to present a mainly data adaptive analysis of the matching function with a minimum of restrictions on the empirical model. Recently developed marginal integration techniques (see e.g. Linton & Nielsen, 1995) allow for nonparametric analysis, which avoids so far necessary restrictive parametric assumptions.

2. Nonparametric Regression Estimation

We consider additive model with smooth functions f_α and interactions $f_{\alpha\beta}$:

$$Y = m(X) + \varepsilon \quad , \tag{8}$$

$$m(x) \;=\; c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{1\le\alpha<\beta\le d} f_{\alpha\beta}(x_\alpha, x_\beta)\;, \tag{9}$$

where $X = (X_1, \ldots, X_d)$ is a vector of explanatory variables, and ε is the disturbance. Apart from the advantage of intepretability, Stone (1985) has proved that such a model does not suffer from the curse of dimensionality. We use the marginal integration approach since for this estimator derivative estimation (Severance-Lossin and Sperlich, 1999), estimation of interaction terms and testing their significance (Sperlich, Tjøstheim and Yang, 1999) is well developed. These tools are extremely useful for our economic analysis. In (9), $\{f_\alpha(\cdot)\}_{\alpha=1}^{d}$ and $\{f_{\alpha\beta}(\cdot)\}_{1\le\alpha<\beta\le d}$ are uniquely identified when fixed in the vertical direction. Let $X_{\underline{\alpha}}$ be the $(d-1)$-dimensional random variable obtained by removing X_α from $X = (X_1, \ldots, X_d)$, and $X_{\underline{\alpha\beta}}$ defined analogously. We denote the (marginal) density of $X_{\underline{\alpha}}$ and of X by $\varphi_{\underline{\alpha}}(x_{\underline{\alpha}})$ and $\varphi(x)$. The marginal effects are defined by

$$F_\alpha(x_\alpha) = \int m(x_\alpha, x_{\underline{\alpha}})\varphi_{\underline{\alpha}}(x_{\underline{\alpha}})dx_{\underline{\alpha}}, \tag{10}$$

$$F_{\alpha\beta}(x_\alpha, x_\beta) = \int m(x_\alpha, x_{\beta,}x_{\underline{\alpha\beta}})\varphi_{\underline{\alpha\beta}}(x_{\underline{\alpha\beta}})dx_{\underline{\alpha\beta}}. \tag{11}$$

F_α corresponds to f_α and $F_{\alpha\beta} - F_\alpha - F_\beta$ to $f_{\alpha\beta}$ up to a constant. Then, estimate

$$\widehat{F}_\alpha(x_\alpha) = \frac{1}{n}\sum_{l=1}^{n} \widehat{m}(x_\alpha, X_{l\underline{\alpha}}) \quad , \quad \widehat{F}_{\alpha\beta}(x_\alpha, x_\beta) = \frac{1}{n}\sum_{l=1}^{n} \widehat{m}(x_\alpha, x_\beta, X_{l\underline{\alpha\beta}}) \tag{12}$$

using kernel smoothers. To compute $\widehat{m}(x_\alpha, X_{l\underline{\alpha}})$ we make use of a specific kind of multidimensional local polynomial kernel estimation; see Ruppert and Wand (1994) for the general case. This allows us to estimate simultaneously the functions and its derivatives ($\widehat{f'_\alpha} = \widehat{F'_\alpha}$) Explicit theorems and proofs can be found in Severance-Lossin & Sperlich (1999) and Sperlich, Tjøstheim and Yang (1999). In the latter are also presented the procedures to test for possible interaction.

3. Data and Economic Model

We first estimated a parametric benchmark model regressing log unemployment to job exits $F_{i,t}$ in labor market district i over period t on log unemployment and vacancies $V_{i,t-1}$. Accounting for the bias arising from differences in size of districts (Münch, Svejnar and Terrell, 1998), we divide all variables by the size of the labor force at the beginning of the month. As in Boeri (1994), we account for a diminishing job finding probability of unemployed at longer spells by allowing different matching efficiencies for long-term $U^*_{i,t-1}$ and newly unemployed $I_{i,t-1}$. Burda & Profit (1996) have demonstrated that residuals of the static Czech matching function show strong serial correlation explained by a time lag between matching and hiring of workers with firms. We account for this by including a lagged dependent variable into the estimation. Finally, we capture the heterogeneity among districts and time trends by introducing corresponding fixed

effects:

$$\ln F_{i,t} = \nu_i + \delta_t + \gamma \ln F_{i,t-1} + \alpha_U \ln U^*_{i,t-1} + \alpha_I \ln I_{i,t-1} + \alpha_V \ln V_{i,t-1} + \epsilon_{i,t}, \quad (13)$$

A most general nonparametric analogue would be the model

$$Y_{it} = \alpha_i + \gamma_t + F(X_{it} + \mu_i + w_t) + \epsilon_{i,t}. \quad (14)$$

Looking for estimators which fulfill the conditions: 1. Identifiability 2. nonparametric estimation of F (having only one observation per district i and time t), and 3. equivalence to the parametric model (13), leads to

$$\begin{aligned} \hat{\alpha}_i &= \bar{y}_{i\cdot} + c_1, \quad \hat{\mu}_i = \bar{x}_{i\cdot} + c_3 \\ \hat{\gamma}_i &= \bar{y}_{\cdot t} + c_2, \quad \hat{w}_i = \bar{x}_{\cdot t} + c_4 \end{aligned}$$

where c_1, c_2, c_3 and c_4. The functional F was decomposed as in (9). Notice that estimating (14) and (8) is equivalent in the nonparametric world.

4. Results and Conclusions

In the parametric estimation we found that the coefficient on (long-term) unemployment is positive and highly significant, the coefficient on vacancies is, except for 1992, positive but very small and insignificant in most years. Moreover, we find a positive and significant coefficient of lagged unemployment inflows, which is however smaller than the coefficient on unemployment stocks. This is at odds with Boeri (1994) who found a higher matching efficiency of newly unemployed. One explanation could be that unemployment inflows of the previous period are an inadequate measure for the short-term unemployed. If newly unemployed find new jobs within the same month, previous month's inflows overestimate short-term unemployment. Comparing regression over time reveals the instability of matching coefficients. This implies that structural changes during the transformation process had a strong impact on unemployment-to-job exits, and alter the districts' fixed effects over time. Therefore, we estimate the matching function on a year-by-year basis. In the article are given nonparametric estimates for all years, for the marginal impact functions, interactions and return to scales. Only for 1992 interaction between long and short term unemployment turned out to be significant. This is in contradiction with the random search hypothesis in economic theory. In general we detected strong non-uniformities in the job-matching process in the Czech Republic during the transition period. For the marginal impacts we found a negatively sloped or hump-shaped marginal contribution of vacancies in some years, which helps to explain why the coefficient on vacancies is small and insignificant in the parametric model. For the return to scales we did density plots for each year. It is demonstrate that the distribution of local returns to scale is skewed to the left with a single mode clearly above one. For 1992, 43% of all observations exhibit increasing returns to scale. In

1993, this fraction increases to 55%, and 82% in 1994 and 1996 (in 1995 it drops to 41%). In 1995 and 1996, the variance of the distribution of local returns to scale increases compared to previous years. Hence, the nonparametric estimates confirm the findings of slightly increasing returns to job-matching on Czech labor markets as in Profit (1997). Moreover, we find some seasonal variation in returns to scale estimates with higher values during spring and summer. This is an important finding, since "local" returns to scale may be responsible for the emergence of multiple equilibria in unemployment rates. The fact that Czech labor market districts with above average unemployment rates have increasing returns to job-matching is consistent with multiple equilibria with these districts being trapped in a *bad* equilibrium. Another important finding is the positive correlation of active labor market policies (program participation, staffing of district labor offices and ALMP expenditures) and the matching technology.

Références

Boeri, T. (1994). "Labour Market Flows and the Persistence of Unemployment in Central and Eastern Europe." In: OECD (ed.), *Unemployment in Transition Countries: Transient or Persistent?*, Paris.

Boeri, T., and M.C. Burda (1996). "Active Labour Market Policies, Job Matching and the Czech Miracle," *European Economic Review*, **40**, 805-817

Burda, M.C. (1994). "Modeling Exits from Unemployment in Eastern Germany: A Matching Function Approach." In: H. König and V. Steiner (eds.), *Arbeitsmarktdynamik und Unternehmensentwicklung in Osteuropa*, Baden-Baden: Nomos.

Burda, M.C., and S. Profit (1996). "Matching Across Space: Evidence on Mobility in the Czech Republic," *Labour Economics*, **3**, 255-278.

Münich, D., J. Svejnar and K. Terrell (1995). "Regional and Skill Mismatch in the Czech and Slovak Republics," In: OECD (ed.), *The Regional Dimension of Unemployment in Transition Countries*, Paris.

Münich, D., J. Svejnar and K. Terrell (1998). "The Worker-Firm Matching in Transition Economies: (Why) Are The Czechs More Successful than Others?" *Preprint.*

Profit, S. (1997). "Twin Peaks in Regional Unemployment and Returns to Scale in Job-Matching in the Czech Republic," *Discussion Paper 63, SFB 373 Berlin.*

Ruppert, D. and M.P. Wand (1994). "Multivariate Locally Weighted Least Squares Regression," *Ann. Statist.*, **22**, 1346-1370.

Severance-Lossin, E. and S. Sperlich (1999). "Estimation of Derivatives for Additive Separable Models," *Statistics*, **33**, 241-265.

Sperlich, S., Tjøstheim, D. and L. Yang (1999). "Nonparametric Estimation and Testing of Interaction in Additive Models," *forthcoming in Econometric Theory.*

Stone, C.J. (1985). "Additive Regression and other Nonparametric Models," *Ann. Statist.*, **13**, 689-705.

Modèle additif de régression sous des conditions de mélange

Christine CAMLONG-VIOT

Laboratoire de Statistique et Probabilités
Université Paul sabatier
Toulouse 31062 Cedex
camlong@cict.fr ou christine.camlong@wanadoo.fr

Soutenance de thèse du 28 Novembre 2000

Résumé

Ce travail aborde le problème de l'estimation non paramétrique de la fonction de régression multivariée.

Dans la pratique, lorsqu'il s'agit d'étudier une fonction de régression, on a souvent affaire à une fonction multivariée c'est-à-dire une fonction où on n'a pas une seule mais plusieurs variables explicatives (la variable explicative est un vecteur de dimension d). Afin de ne pas imposer de forme *a priori* à la fonction à estimer on préférera se placer dans le cadre plus général des modèles non paramétriques. Toutefois, comme sous ces modèles les résultats que l'on peut obtenir subissent l'influence de la dimension de la variable explicative, on restreindra légèrement le modèle en considérant les modèles non paramétriques additifs sous lesquels il s'agira non plus d'estimer une seule fonction multivariée mais plusieurs fonctions univariées. Sous ces conditions, on utilisera une des méthodes d'estimation les plus récentes : la méthode de l'intégration marginale qui nous permettra d'obtenir des estimateurs de chaque fonction univariée et par la suite un estimateur de la fonction de régression multivariée.

Ensuite un problème apparaît : vaut-il mieux restreindre le modèle en considérant que la fonction de régression est additive pour avoir une meilleure vitesse de convergence de l'estimateur ou bien vaut-il mieux avoir une moins bonne vitesse avec un modèle plus général ? Dans l'optique de tester l'additivité de la fonction de régression étudiée, on propose une statistique de test pour laquelle, grâce aux résultats de convergence obtenus pour l'estimateur de la fonction de régression construit par la méthode de l'intégration marginale, on montre la normalité

asymptotique. Afin de pouvoir étendre les résultats obtenus avec cette méthode d'estimation à de la prévision en séries chronologiques, on considérera de plus que les observations utilisées pour faire l'estimation ne sont pas indépendantes mais asymptotiquement indépendantes.

Références

Camlong, Ch. (1999). Convergence presque-sûre de l'estimateur à noyau d'une fonction de régression additive sous une hypothèse de mélange. *C. R. Acad. Sci. Paris, Série I*, **329**, 75-78.

Camlong, Ch., Sarda, P. and Vieu, Ph. (2000). Additive time series: the kernel integration method. *Math. Methods Statist.*, (en cours d'impression).

Camlong, Ch. (2000). *Modèle additif de régression sous des conditions de mélange.* Synthèse de travaux de recherches présentés en vue de l'obtention du Doctorat de 3^{me} cycle de l'Université P. Sabatier, Toulouse 3, Laboratoire de Statistique et Probabilités le 28 Novembre 2000.

Camlong, Ch. (2000). Testing additivity in nonparametric regression under mixing conditions. *Preprint.*

Contributions à la Statistique Multidimensionnelle Opératorielle

Yves ROMAIN

Laboratoire de Statistique et Probabilités, Université Paul Sabatier
romain@cict.fr

Soutenance d' Habilitation du 11 Décembre 2000

Mots-clés : analyse en composantes principales, fonction aléatoire hilbertienne, analyse canonique, analyse canonique relative, sous-espaces, opérateur de covariance, loi asymptotique de valeurs, vecteurs et projecteurs propres, projection orthogonale, mesure d'association, principe d'incertitude, mesure aléatoire, mesure spectrale, opérateurs autoadjoints, statistique quantique ou non commutative, algèbre de von Neumann, ampliation fonctionnelle, produit, somme et différence tensoriels fonctionnels.

Résumé

La première partie est consacrée à la Statistique factorielle considérée dans un cadre fonctionnel et aux études asymptotiques associées lors d'approximation par échantillonnage i.i.d.. Plus précisément, nous avons d'abord étudié l'Analyse en Composantes Principales (ACP) d'une fonction aléatoire hilbertienne et ses approximations par échantillonnage. Sont alors obtenus des résultats généraux de convergence, presque sûre et en loi, concernant les suites des éléments spectraux (valeurs, vecteurs, projecteurs propres) des opérateurs de covariance dont les décompositions spectrales fournissent l'analyse échantillonnée. La méthodologie statistique et les techniques d'obtentions des lois asymptotiques (qui, sous hypothèses elliptiques, sont explicitées dans un langage tensoriel) sont alors transportables aux autres analyses factorielles et à diverses méthodes dérivées (modulo leurs spécificités bien sûr). Nous obtenons ainsi, en dimension infinie, un cadre synthétique d'étude de comportement asymptotique des méthodes de Statistique multidimensionnelle qui se ramènent à l'analyse spectrale d'un opérateur autoadjoint compact positif. Outre l'ACP (centrée ou non) déjà citée, nous avons plus particulièrement travaillé sur l'ACP réduite (les trois réductions possibles sont considérées), sur l'Analyse Canonique de sous-espaces factoriels (qui permet la comparaison de deux populations) et, récemment, sur l'Analyse Canonique de deux sous-espaces fermés relativement à un troisième (dont le coefficient maximal englobe plusieurs "mesures d'association" connues).

Dans la deuxième partie, notre intérêt se focalise sur l'environnement opératoriel stochastique de méthodes multidimensionnelles et sur l'argumentation, même en dimension finie, d'une approche "fonctionnelle". À partir d'un problème

d'ordre "linguistique" (comment comparer des explicitations matricielles à des écritures fonctionnelles ?), nous avons d'abord élaboré un dictionnaire matriciel-opératoriel dans un cadre réel euclidien. Puis, nous avons tenté de dégager les outils et notions clés d'une telle approche afin de mieux situer le champ de nos études dans la littérature statistique et au delà. La projection orthogonale, en tant que notion pivotale, et les algèbres de Von Neumann, en tant que structure fondamentale (elles sont engendrées par les projections qu'elles contiennent), nous sont apparues essentielles et un travail particulier a été effectué autour des projecteurs orthogonaux et des opérateurs associés utiles en Statistique multidimensionnelle.

Dans la troisième partie, les travaux portent sur des domaines connexes à nos thèmes habituels. Ils sont, pour nous, des conséquences générées par l'approche préconisée précédemment. Un premier exemple est celui de l'étude d'une "algèbre" engendrée par des opérateurs tensoriels, nommés "ampliations fonctionnelles" et qui permettent de plonger dans un espace commun plusieurs opérateurs, éventuellement aléatoires, définis dans des espaces différents. Cette étude sur des éléments tensoriels fonctionnels (dont la "maniabilité technique" est à souligner, ce qui fournit divers résultats d'ordre calculatoire, spectral, de perturbation...) est motivée, d'une part, par l'étude des extensions tensorielles des analyses (factorielles) d'opérateurs, d'autre part, par le caractère intrinsèquement opératoriel et tensoriel de la Statistique quantique. Un second exemple est une première approche pour considérer, d'un point de vue statistique, les deux types de principes d'incertitude les plus répandus : pour le premier type, nous étudions l'inégalité de Cramer-Rao dans ses versions classique et quantique, alors que, pour le second type, nous proposons une interprétation de la relation d'incertitude en terme d'une analyse canonique complexe de deux sous-espaces bien choisis. Un troisième exemple concerne les liens entre les produits (tensoriel et de convolution) de mesures (spectrales et aléatoires) et les ampliations fonctionnelles. Ce travail a donné des développements récents concernant la convolée de deux mesures aléatoires qui s'appliquent aux processus (hilbertiens) stationnaires; il est intéressant de noter que la mesure aléatoire associée à cette convolée est celle du produit tensoriel des deux processus stationnaires considérés.

En conclusion, la réflexion menée et les résultats obtenus dans les parties précédentes nous confirment la place particulière, tant d'un point de vue théorique qu'appliqué, que devrait jouer la "tensorisation" (opératorielle, stochastique, réelle et complexe) de la Statistique multidimensionnelle dans nos futurs travaux.

Référence

Romain, Y. (2000). *Contributions à la Statistique Multidimensionnelle Opératorielle.* Synthèse de travaux de recherches présentés en vue de l'obtention de l'Habilitation à diriger des recherches à l'Université P. Sabatier, Toulouse 3, Laboratoire de Statistique et Probabilités le 11 Décembre 2000.

Contributions à l'Estimation Fonctionnelle

Pascal SARDA

Laboratoire de Statistique et Probabilités
Université Paul Sabatier
Toulouse 31062 Cedex
sarda@cict.fr

Soutenance d' Habilitation du 11 Décembre 2000

Mots-clés : *Estimation fonctionnelle, Estimateurs à noyau, Splines de régression, Régression linéaire locale, Régression, Densité, Fonction de répartition, Fonction de hasard, Quantiles conditionnels, Séries chronologiques, Prévision, Discontinuité, Validation croisée, Composantes Principales Additives, Modèle linéaire fonctionnel.*

Résumé

Les travaux présentés ont trait à l'estimation fonctionnelle, domaine de la statistique qui suscite un intérêt croissant depuis une quarantaine d'années. Plusieurs aspects sont étudiés dans ce cadre général. Tout d'abord nous établissons des propriétés asymptotiques d'estimateurs non paramétriques (essentiellement les estimateurs à noyau) de différents paramètres fonctionnels (parmi lesquels régression, densité, fonction de hasard, densité conditionnelle). Ces travaux sont pour la plupart réalisés dans le cadre d'estimateurs construits à partir d'observations non indépendantes et englobent la convergence uniforme presque complète ainsi que les vitesses de convergence.

Lorsqu'on s'intéresse à la mise en œuvre d'estimateurs à noyau, un problème crucial est celui du choix du paramètre de lissage (largeur de fenêtre). Il a naturellement été au premier plan dans la littérature pendant une quinzaine d'années et beaucoup d'articles traitent encore de ce sujet. Pour notre part, nous avons étudié des méthodes de sélection de largeurs de fenêtre reposant sur la validation croisée et en particulier des critères de sélection de fenêtres locales. Nous nous sommes attachés à montrer l'optimalité asymptotique de ces méthodes.

Le "fléau de la dimension", qui se traduit par la dégradation des vitesses de convergence lorsque la dimension augmente, concerne la plupart des estimateurs non paramétriques de paramètres fonctionnels et est depuis une vingtaine d'années un autre domaine d'intérêt en estimation fonctionnelle. Notre propre contribution à ce problème comprend deux volets. Nous avons étudié l'estimation de composantes principales additives définies en minimisant la variance de fonctions additives du vecteur des observations sous des conditions d'orthogonalité.

Les estimateurs proposés reposent sur les splines de régression. L'existence, l'unicité et la convergence L^2 de ces estimateurs ont été établies. Nous avons également étudié le modèle additif de régression dans le cas d'observations non indépendantes : l'estimateur est ici construit à l'aide d'une intégration marginale d'un estimateur à noyau.

Dans de nombreux problèmes de régression, la variable explicative est une courbe (modélisée par un élément d'un espace de Hilbert) et la variable à expliquer un scalaire. La littérature sur ces problèmes est relativement ancienne bien que peu abondante quant à l'approche fonctionnelle. Les méthodes utilisées sont principalement des adaptations du modèle linéaire et présentent certaines limites. L'intérêt de méthodes "fonctionnelles" a été mis en lumière dans des travaux récents (ceux de Ramsay et Silverman notamment) mais peu de résultats théoriques sont à ce jour disponibles : de nombreux problèmes, ouvrant un champ très large d'un point de vue théorique et appliqué, devraient motiver des recherches futures. Dans notre travail, nous nous sommes intéressés au modèle linéaire fonctionnel pour lequel l'espérance conditionnelle est modélisée par un opérateur linéaire continu Ψ ou de manière alternative par le produit scalaire entre la courbe prédictive et un coefficient fonctionnel α. Dans ce cadre, des propriétés (existence, unicité, convergence) de deux estimateurs de l'opérateur Ψ et/ou de la fonction α ont été établis. Le premier repose sur une régression sur composantes principales fonctionnelles et le second sur l'approximation du coefficient fonctionnel par une fonction spline. Nous nous sommes intéressés ensuite au test de la nullité de l'opérateur Ψ (ou de celle de α), test pour lequel deux statistiques ont été proposées. Ces études ont également donné lieu à un travail appliqué au travers de simulations et d'exemples réels.

Enfin, d'autres travaux concernant l'estimation de points de discontinuité de l'intensité d'un processus de Poisson non homogène ou encore l'estimation de quantiles conditionnels ont été réalisés et devraient également être prolongés dans le futur.

Référence

Sarda, P. (2000). *Contributions à l'Estimation Fonctionnelle.* Synthèse de travaux de recherches présentés en vue de l'obtention de l'Habilitation à diriger des recherches, à l'Université P. Sabatier, Toulouse 3, Laboratoire de Statistique et Probabilités le 11 Décembre 2000.

A propos des flux paramétriques

Jim RAMSAY

Département de Psychologie
1205 Avenue du Dr. Penfield
Montreal, H3A 1B1, Quebec
ramsay@psych.mcgill.ca

Exposé du 12 Décembre 2000

Résumé

Doit-on toujours forcément choisir entre la valeur du paramètre spécifiée par l'hypothèse nulle et une contre-hypothèse composite ? Pas du tout. En régression ridge, par exemple, ou dans les problèmes de sélection de fenêtre en régression non-paramétrique, on dispose d'un continuum de modèles ajustés aux données. Dans de telles situations, il est généralement possible d'améliorer l'estimation des paramètres en faisant un compromis entre le modèle général et le modèle de dimension réduite précisé par l'hypothèse nulle.

Un flux paramétrique se définit comme une trajectoire liant une estimation de paramètre de faible dimension à une contre-hypothèse multidimensionnelle le long d'une courbe déterminée par un critère d'ajustement. Le flux est caractérisé par une équation différentielle ordinaire qui, dans bien des cas, est à la fois facile à formuler et à résoudre numériquement.

Dans cet exposé, j'illustrerai concrètement le concept de flux paramétrique dans différentes situations. Comme nous le verrons, il est souvent possible d'améliorer de façon substantielle l'estimation de paramètres en choisissant judicieusement des points le long du flux. C'est notamment le cas lorsque l'échantillon est de petite taille ou que le rapport signal-bruit est faible, parce que la variance expérimentale a alors tendance à croître de façon marquée au voisinage de l'optimum, tandis que le biais décroît presque linéairement. Nous verrons aussi qu'en faisant appel à la validation croisée, il est souvent possible d'estimer un point du flux paramétrique qui améliore sensiblement l'erreur quadratique moyenne.

Références

Chambers, J. M. and Hastie, T. J. (1991) *Statistical Models in S.* London: Chapman & Hall.

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical society, Series B.* **45**, 311-354.

James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium, Vol. 1.* 361-379.

Ramsay, J.O. (1970). A family of gradient methods for optimization. *The Computer Journal,* **13**, 413-417.

Nonlinear alignment of time series with applications to varve chronologies

Dag TJOSTHEIM

Département de Mathématiques
Université de Bergen
5007, Bergen, Norvège
dagt@skolem.mi.uib.no

Exposé du 12 Décembre 2000

Abstract

The problem of aligning time series arises naturally in a number of fields. The results of the talk are of a general nature, but the motivation comes from the problem of extracting paleoclimatic signals from varve chronologies. Glacial varves are laminated sediments that record annual depositional cycles in certain lakes that are fed by glaciers. In this sense varve thicknesses are an indication of yearly temperature changes (much as for tree rings). An important aspect of the varve problem is to align series from different locations. Traditionally, alignment problems have been treated using linear methods such as coherency and cross correlation. I will look at nonlinear methods derived from recent work on nonlinear dependence modeling.

Références

B. H. Auestad, R.H. Shumway, D. Tjostheim and K.L. Verosub. (2000). Nonlinear alignment of time series with applications to varve chronologies. Preprint.
Skaug, H. and Tjostheim, D. (1993). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika*, **80**, 591-602.
Tjostheim, D. (1996). Measuring of dependence and tests of independence. *Statistics*, (invited paper), **28**, 249-280.

Boosting Wavelets in electrophoresis

Ja-Yong KOO *
En collaboration avec Hervé CARDOT et Alan TRUBUIL

* Adresse pour correspondance:
Department of Statistics
Hallym University
Chunchon, Kangwon-Do 200-702
jykoo@sun.hallym.ac.kr

Exposé du 15 Janvier 2001

Abstract

Electrophoresis is a biochemical process widely used in life sciences and genetics. Recovering information on molecules locations and proportions from electrophoresis images may be viewed as a linear inverse problem in the context of a regression setup. The target function to be estimated is known to be positive and spatially inhomogeneous. In order to recover such functions, wavelet shrinkage and boosting ideas are used. The method is illustrated on simulated data and applied to electrophoresis.

The deepest regression method

Peter ROUSSEUW

Department of Mathematics and Computer Science
Universitaire Instelling Antwerpen
Universiteitsplein 1, B-2610 Antwerp, Belgique
rousse@uia.ua.ac.be
http://win-www.uia.ac.be/u/statis

Exposé du 22 Janvier 2001

Abstract

Deepest regression (DR) is a method for linear regression introduced by Rousseeuw and Hubert (1999). The DR method is defined as the fit with largest regression depth relative to the data. We will show that DR is a robust method, with breakdown value that converges almost surely to 1/3 in any dimension. We construct an approximate algorithm for fast computation of the DR in more than two dimensions. From the distribution of the regression depth we derive tests for the true unknown parameters in the linear regression model. Moreover, we construct confidence regions based on bootstrapped estimates. For bivariate datasets we use the maximal regression depth to construct a test for linearity versus convexity/concavity. We also extend the DR to polynomial regression. Finally, DR is applied to the Michaelis-Menten model of enzyme kinetics, where it resolves a long-standing ambiguity.

Keywords: Algorithm, Applications, Inference, Linearity.

References

Rousseeuw, P.J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association*, **94**, 388-402.
Van Aelst, S., Rousseeuw, P.J., Hubert, M. and Struyf, A. (2001), "The Deepest Regression Method," *Journal of Multivariate Analysis*, to appear.

Estimations de l'occupation des sols a partir de l'evolution temporelle des images du capteur vegetation de SPOT

Robert FAIVRE
En collaboration avec Hervé CARDOT, Michel GOULARD
et Héloïse VIALARD

Adresse pour correspondance:
INRA Toulouse, Biométrie et Intelligence Artificielle
31326 Castanet-Tolosan cedex, France
faivre@toulouse.inra.fr

Exposé du 5 Février 2001

Mots-clés : désagrégation, données longitudinales, régression sous contrainte, lissage, B-splines, modèle multilogit, analyse en composantes principales fonctionnelle, capteur SPOT 4/ Végétation

---

Résumé

Le capteur Végétation du satellite SPOT4 fournit de manière quasi-quotidienne (haute répétitivité temporelle) des images de l'Europe à faible résolution spatiale, chaque pixel représente une zone de 1km $\times$ 1km. Les informations fournies par ce capteur dans le visible, le proche et le moyen infra-rouge, permettent de caractériser l'état de développement de la végétation à l'échelle d'une région (Tucker, 1979). Les pixels observés, appelés *mixtes* (Faivre et Fisher, 1997), mélangent différentes informations au sens où ils représentent différentes parcelles agricoles, dont la taille en France est nettement inférieure au km^2, et donc différentes cultures ou thèmes d'intérêt (blé, maïs, orge, forêt...).

Dans le cadre de pixels mixtes, la réflectance observée sur un pixel est un mélange des réflectances de chaque thème cultural. Dans le visible, on suppose de plus que ce mélange est linéaire, c'est-à-dire que la réflectance $Y_i(t)$ d'un pixel i à une date t, est une combinaison linéaire des réflectances $\rho_{ij}(t)$ de chaque classe

j pondérée par les proportions π_{ij} :

$$Y_i(t) = \sum_{j=1}^{p} \pi_{ij}\rho_{ij}(t) + \varepsilon_i(t)$$

à une erreur $\varepsilon_i(t)$ près. On suppose que sur la période d'intérêt, l'occupation des sols est fixe, c'est-à-dire que les pratiques culturales (choix de la culture) sont établies.

Deux questions se posent.

1. Comment retrouver les réflectances $\rho_{ij}(t)$ connaissant les π_{ij} ?
2. Comment retrouver les occupations π_{ij} connaissant les $\rho_{ij}(t)$?

Faivre et Fischer (1997) ont proposé une approche pour répondre à la première question. Il s'agissait, pour chaque date d'observation et pour chaque longueur d'onde, d'estimer la distribution sur l'image de la réflectance associée à chaque type de culture, connaissant le plan d'occupation des sols et de prédire localement les variations de réflectances pour une même classe. Pour cela, un modèle linéaire mixte a été utilisé. Cette approche a été poursuivie par Faivre, Delécolle et Guérif (1999) pour estimer les évolutions des réflectances de chaque classe sur chaque pixel, la désagrégation étant réalisée indépendamment pour chaque date. La dynamique des réflectances ainsi prédite pour chaque classe sur chaque pixel a ensuite été utilisée par assimilation dans un modèle de croissance d'une culture pour prédire la production régionale en blé d'hiver (Husson et Faivre, 2000).

La seconde question aborde un problème de classification. Etant donnée la résolution des images (pixel de 1 km x 1 km), il ne s'agit pas ici d'affecter à chaque pixel une classe d'occupation mais d'évaluer le vecteur des pourcentages occupés par chaque thème cultural. Deux sources de données sont couramment utilisées pour résoudre le problème d'évaluation lié aux pixels mixtes. La première est d'utiliser les multiples longueurs d'onde du capteur. Malheureusement, le capteur Végétation ne fournit les informations que pour 4 bandes spectrales dans le visible, le proche et moyen infra-rouge. La seconde source est l'information apportée par l'évolution temporelle des réflectances sur chaque pixel.

Notre travail porte sur une réponse à la deuxième question, c'est-à-dire estimer une occupation des sols dans le cadre d'observation de pixels mixtes. Notre objectif est de tenir compte de l'évolution temporelle des pixels sur les images pour séparer les différents types de culture et déterminer le plan d'occupation des sols (POS) de la région c'est-à-dire la proportion de chaque thème cultural à l'intérieur de chaque pixel de 1 km^2. Pour cela, nous disposons d'une petite région sur laquelle nous avons accès à un plan d'occupation des sols. Il s'agit alors d'étendre cette connaissance de l'occupation des sols aux régions avoisinantes afin d'en déduire une prédiction de production à l'aide d'un couplage avec un modèle de croissance et des données météorologiques.

Le problème statistique est la prédiction de proportions à partir d'observations longitudinales. Pour cela, nous avons considéré deux approches.

La première approche est une approche inverse. Elle consiste dans un premier temps à estimer les courbes caractéristiques de réflectance associées à chaque classe du POS. On utilise pour cela les *varying-time regression models* (Hastie & Tibshirani 1993, Hoover *et al.* 1998). Une fois ces courbes estimées, nous prédisons le plan d'occupation des sols d'une zone voisine à l'aide d'un modèle linéaire avec contraintes sur les paramètres qui sont dans cette étape les proportions des thèmes culturaux.

La deuxième approche est directe. Elle repose sur la modélisation des proportions par un modèle linéaire généralisé (Mc Cullagh et Nelder, 1989) et consiste à estimer les paramètres d'une loi multinomiale à l'aide de la fonction de lien multilogit. Les variables explicatives sont les valeurs de la courbe de réflectance du pixel aux différents instants de mesure. Le grand nombre de variables explicatives nous conduit à "stabiliser" l'estimateur au moyen d'une réduction de la dimension du problème. Dans le cadre de la régression linéaire fonctionnelle, Hastie et Mallows (1993) et Cardot *et al.* (1999) ont proposé une stabilisation de l'estimateur au moyen d'une analyse en composantes principales fonctionnelle (Deville 1974, Ramsay et Silverman, 1997) des courbes explicatives. Nous proposons d'appliquer cette méthode dans le cadre du modèle multilogit. Compte-tenu du nombre important de fonctions de régression à estimer dans notre exemple, neuf si le nombre de classe du POS est de dix, une approche par pénalisation (Marx and Eilers, 1999) ne semble pas réaliste puisqu'il faudrait déterminer les valeurs des 9 paramètres de lissages par validation croisée.

La comparaison est effectuée, pour chaque bande spectrale, sur un site pilote sur lequel nous disposons de l'occupation des sols obtenue par classification supervisée d'images Spot à haute résolution spatiale (pixel de 20m x 20m). On évalue ainsi la capacité de chaque canal à prédire l'occupation des sols.

Le plan de l'exposé est le suivant. Après une présentation des données, nous décrivons dans les deux parties suivantes l'approche par courbes caractéristiques et le modèle multilogit. Ces méthodes sont ensuite comparées sur la base des estimations (prédictions) du POS, pour un échantillon test de pixels, obtenues en fonction de l'évolution temporelle observée par le capteur Végétation du satellite SPOT 4. L'approche de type multilogit donne généralement de meilleurs résultats et l'utilisation d'indices de végétation composites est plus efficace.

Les procédures d'estimation en langage Matlab ainsi qu'u rapport technique détaillé (Cardot *et al.* 2000) sont disponibles sur demande auprès des auteurs.

Références

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional Linear Model. *Statist.*

& Prob. Letters, **45**, 11-22.
Cardot, H., Faivre, R. et Goulard, M. (2000). Estimation de l'occupation de sols à partir de l'évolution temporelle des images du capteur Végétation de SPOT. *Rapport Technique, INRA Toulouse, Biométrie et Intelligence Artificielle*, **2**, 35 p.
Deville, J.C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**, 3-97.
Faivre R., R. Delécolle and M. Guérif, (1999). Synthetic Mapping of Crop Dynamics with Pseudo-Vegetation Data. ALPS'99 : International Conference and Workshops, WK3 "Remote Sensing and vegetation productivity", CNES(éd.), Méribel 18-22 janvier 1999, **O**-10.
Hastie, T.J. and Tibshirani, R.J, (1993). Varying-Coefficient Models (with discussion). *J. Roy. Statist. Soc., B*, **55**, 757-796.
Hastie, T.J. and Mallows, C., (1993). A discussion of "A statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman. *Technometrics*, **35**, 140-143.
Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.
Husson A. and Faivre R., (2000). Integration of VEGETATION and HRVIR data into yield estimation approach. *Congrès Végétation 2000*, Belgirate 3-6 avril, Italie.
Marx, B.D. and Eilers P.H. (1999). Generalized Linear Regression on Sampled Signals and Curves: A *P*-Spline Approach. *Technometrics*, **41**, 1-13.
McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall.
Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag.
Richardson A.J. and Wiegang, C.L., (1977). Distinguishing Vegetation from Soil Background Information. *Photogrammetric Engineering and Remote Sensing*, **43**, 1541-1552.
Tucker, C.J., (1979). Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, **8**, 127-150.

Estimation pour le modèle de Lotka-Volterra

Sorana FRODA

Univ du Quebec a Montreal
froda@math.uqam.ca, froda@cict.fr

Exposé du 26 Février 2001

Résumé

Le modèle déterministe de Lotka-Volterra date des années 1930 et décrit de façon simple l'interaction entre prédateur et proie, quand la proie est la principale source de nourriture du prédateur. Nous proposons une (autre) version aléatoire de ce modèle et une méthode d'estimation qui permet de prédire les tailles des populations animales. Notre méthode exploite quelques propriétés qualitatives du système de Lotka-Volterra et elle est appliquée à deux ensembles de données "historiques" (vison–rat musqué et lynx–lièvre).

Références

Colavita, G. and Froda, S. (2000) Estimating the parameters of the Lotka-Volterra system of equations. Preprint.
Haberman, R. (1977). Mathematical models, Prentice-Hall.
Renshaw, E. (1991). Modelling biological populations in space and time, Cambridge Univ. Press.

Perturbations d'operateurs aleatoires et applications

Jeane FINE

LSP et IUFM, Toulouse
fine@cict.fr

Exposé du 5 Mars 2001

1. Résumé

La théorie des perturbations pour les opérateurs linéaires présentée par Kato (1966, 1980) a été adaptée (Fine, 1981, 1987) à un problème particulier utile en statistique multidimensionnelle. Nous rappelons ces résultats, qui s'appliquent directement à l'Analyse en Composantes Principales, et montrons, dans la dernière partie, des applications en Analyse Canonique.

2. Un problème de perturbation; les résultats de Kato

Soit ε un réel de $]0\,;\,1[$, V et U des opérateurs compacts autoadjoints et positifs d'un espace de Hilbert séparable E, et $V(\varepsilon)$ l'opérateur défini par $V(\varepsilon) = V + \varepsilon U$. Il s'agit d'un problème de perturbation, V est l'opérateur non perturbé, $V(\varepsilon)$ l'opérateur perturbé et εU la perturbation. L'objectif est d'écrire les éléments propres (projecteurs propres, valeurs propres, vecteurs propres) de $V(\varepsilon)$ en fonction de ceux de V et de U et de puissance de ε. Kato (1966) a apporté la solution à ce problème et l'a généralisée au problème de perturbation : $V(\varepsilon) = V + \varepsilon U(\varepsilon)$ avec $V(\varepsilon)$ différentiable à l'origine (perturbabion asymptotique) ou bien avec $U(\varepsilon)$ s'écrivant comme une série de Taylor à l'origine (perturbation analytique). Dans les deux cas, notons que $(U(\varepsilon))_{\varepsilon\in]0,1[}$ converge vers un opérateur U quand ε tend vers 0. Les résultats obtenus sont les suivants. Soit λ une valeur propre d'ordre k de V, P le projecteur propre associé et S l'opérateur défini par l'inverse généralisée de $V-\lambda P$, soit: $S = (V-\lambda P)^{-}$. Alors, il existe $\varepsilon_0 \in]0,1[$, tel que pour $\varepsilon < \varepsilon_0$, il existe exactement k valeurs propres de $V(\varepsilon)$ dans l'intervalle $]\lambda - \varepsilon_0; \lambda + \varepsilon_0[$, notées $(\lambda_i(\varepsilon))_{i=1,\ldots,k}$ et rangées dans l'ordre décroissant. On note $P(\varepsilon)$ la somme des projecteurs propres associés à ces k valeurs propres. On peut alors écrire:

$$P(\varepsilon) = P - \varepsilon(PU(\varepsilon)S + SU(\varepsilon)P) + O(\varepsilon^2),$$

$$\lambda_i(\varepsilon) = \lambda + \varepsilon\mu_i(\varepsilon) + O(\varepsilon^2)$$

où $\mu_i(\varepsilon)$ est la *ième* valeur propre dans l'ordre décroissant de l'opérateur $PU(\varepsilon)P$

de rang k, si λ est simple et si on note h un vecteur propre unitaire associé, à condition de bien choisir le vecteur propre unitaire de $V(\varepsilon)$ associé à $\lambda(\varepsilon)$:

$$\lambda(\varepsilon) = \lambda + \varepsilon\mu(\varepsilon) + O(\varepsilon^2)$$

où $\mu(\varepsilon)$ est l'unique valeur propre de l'opérateur $PU(\varepsilon)P$ de rang 1 et

$$h(\varepsilon) = h - \varepsilon SU(\varepsilon)h + O(\varepsilon^2).$$

3. Un problème de perturbation en statistique asymptotique multidimensionnelle

Exemple de l'Analyse en Composantes Principales (ACP) Soit E un espace euclidien et X une v.a. d'ordre 4, définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ à valeurs dans $(E, \mathcal{B}(E))$, E muni de la tribu de ses boréliens. On pose $\mu = \mathbb{E}(X)$ et $V = \mathbb{E}((X-\mu)\otimes(X-\mu))$, espérance et opérateur de covariance de X. Une définition succincte de l'ACP de X est la recherche de la v.a.e. C_1, combinaison linéaire "normée" des composantes de X, $C_1 = \langle x_1, X\rangle_E$ avec $\|x_1\|_E = 1$, de variance λ_1 maximale, puis la recherche de la variable C_2, orthogonale à C_1, de variance λ_2 maximale, et ainsi de suite. Les variables C_i, sont appelées les composantes principales, les vecteurs x_i, les vecteurs principaux, les variances λ_i les valeurs principales. L'Analyse en Composantes Principales (ACP) de X, appelée ACP théorique, s'obtient à partir de la diagonalisation de V : les valeurs propres et vecteurs propres associés fournissent les valeurs principales et vecteurs principaux de l'ACP. A partir d'une suite $(X_i)_{i\in\mathbb{N}^*}$ i.i.d. comme X, on pose : $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ et on définit l'ACP d'échantillonnage ; elle s'obtient à partir de la diagonalisation de V_n. Par définition, l'ACP est stable par échantillonnage si les différents éléments de l'ACP d'échantillonnage convergent vers les éléments correspondants de l'ACP théorique.

Les théorèmes limites Soit $\sigma_2(E)$ l'espace des opérateurs de Hilbert-Schmidt de E muni du produit scalaire de Hilbert-Shmidt. On pose : $U_n = \sqrt{n}(V_n - V)$. Les théorèmes limites dans $\sigma_2(E)$ permettent d'écrire que (V_n) converge p.s. vers V (loi forte des grands nombres), que (U_n) converge en loi vers une gaussienne centrée d'opérateur de covariance celui de $(X-\mu)\otimes(X-\mu)$ (théorème de la limite centrale) et que $(\frac{\|U_n\|}{\sqrt{2\ln\ln n}})$ est p.s. majoré par 1 à partir d'un certain rang (loi du log itéré). En posant : $V_n = V + \frac{1}{\sqrt{n}}U_n$, on est en présence d'un problème de perturbation mais la suite (U_n) est aléatoire et non bornée. Les résultats de la théorie des perturbations de Kato ne peuvent pas être appliqués directement, contrairement à ce qui est pratiqué couramment dans les années soixante et soixante-dix. Une première justification est apportée par Bhattacharya et Gosh en 1978 pour la convergence en probabilité. Il nous a paru utile d'obtenir le même type de développement presque sûrement, ce qui est possible en utilisant la loi du

log itéré. On pose $\varepsilon(n) = \sqrt{\frac{2 \ln \ln n}{n}}$ et $U'_n = \frac{1}{\sqrt{2 \ln \ln n}} U_n$; on a alors, pour presque tout ω,

$$V_n(\omega) = V + \varepsilon(n) U'_n(\omega)$$

et $(U'_n(\omega))$ est une suite non aléatoire bornée par 1 à partir d'un certain rang. Il est alors possible d'adapter la démonstration de Kato à ce problème de perturbation (Fine, 1981, 1987) ; on détaille ci-après les résultats.

4. Premiers résultats en ACP

Nous avons présenté le problème statistique sur l'étude asymptotique de l'ACP. C'est le cadre le plus général: (V_n) est une suite d'opérateurs symétriques positifs convergeant p.s. vers V et (U_n), $U_n = \sqrt{n}(V_n - V)$, est une suite convergeant en loi vers une gaussienne centrée. Ces résultats seront également utilisés pour l'étude asymptotique d'autres méthodes factorielles (Analyse Canonique, Analyse Factorielle Discriminante, Analyse Factorielle des Correspondances, par exemple) ou pour obtenir les propriétés asymptotiques des estimateurs de paramètres de modèles obtenus par diagonalisation (modèles à effets fixes ou à effets aléatoires, par exemple, dans lesquels un échantillonnage non i.i.d. est utilisé).

Diagonalisation de V On note $(\lambda_i)_{i \in I}$ la suite pleine décroissante des valeurs propres de V (c'est-à-dire que l'on écrit les v.p. autant de fois que leurs ordres de multiplicité l'indiquent) et $(h_i)_{i \in I}$ une suite de vecteurs propres unitaires associés. Soit $J \subset I$ tel que $(\lambda_j)_{j \in J}$ soit la suite strictement décroissante des valeurs propres de V. Pour $j \in J$ soit $I_j = \{i \in I \ / \ \lambda_i = \lambda_j\}$, $k_j = card(I_j)$, $P_j = \sum_{i \in I_j} h_i \otimes h_i$ le projecteur propre associé à λ_j et S_j l'opérateur défini par:

$$S_j = (V - \lambda_j P_j)^- = \sum_{k \in J - \{j\}} \frac{1}{\lambda_k - \lambda_j} P_k.$$

On a alors :

$$V = \sum_{i \in I} \lambda_i h_i \otimes h_i = \sum_{j \in J} \lambda_j P_j,$$

la seconde décomposition étant unique.

Diagonalisation de V_n On note $(\lambda_i^n)_{i \in I}$ la suite pleine décroissante des valeurs propres de V_n et $(h_i^n)_{i \in I}$ une suite de vecteurs propres unitaires associés. Pour $j \in J$ soit $P_j^n = \sum_{i \in I_j} h_i^n \otimes h_i^n$ la somme des projecteurs propres associés aux λ_i^n pour $i \in I_j$. On a la propriété suivante : si r désigne la demi distance minimale entre deux valeurs propres distinctes de V, alors, pour n tel que $\varepsilon(n) < r$ et pour tout $j \in J$, il existe exactement k_j v.p. de V_n, les v.p. $(\lambda_i^n)_{i \in J}$,dans l'intervalle $]\lambda_j - r, \lambda_j + r[$.

Résultats A partir d'un certain rang, on a donc, pour $j \in J$:

$$P_j^n(\omega) = P_j - \frac{1}{\sqrt{n}}(P_j U_n(\omega) S_j + S_j U_n(\omega) P_j) + O(\varepsilon^2(n))$$

$$\lambda_i^n(\omega) = \lambda_j + \frac{1}{\sqrt{n}} \mu_i^n(\omega) + O(\varepsilon^2(n))$$

où $\mu_i^n(\omega)$ est la *ième* valeur propre dans l'ordre décroissant de l'opérateur $P_j U_n(\omega) P_j$ de rang k_j, si λ_j est simple et si on note h_j un vecteur propre unitaire associé, à condition de correctement choisir le vecteur propre unitaire de $V_n(\omega)$ associé à $\lambda_j^n(\omega)$, on peut écrire :

$$\lambda_j^n(\omega) = \lambda_j + \frac{1}{\sqrt{n}}\mu_j^n(\omega) + O(\varepsilon^2(n))$$

où $\mu_j^n(\omega)$ est l'unique valeur propre de l'opérateur $P_j U_n(\omega) P_j$ de rang 1 et

$$h_j^n(\omega) = h_j - \frac{1}{\sqrt{n}} S_j U_n(\omega) h_j + O(\varepsilon^2(n)),$$

avec $\varepsilon(n) = \sqrt{\frac{2 \ln \ln n}{n}}$.

On peut, si nécessaire, écrire les développements conjoints de plusieurs éléments propres et pousser les développements à un ordre plus élevé. On peut aussi en déduire les développements de fonctions plusieurs fois différentiables de ces éléments propres en utilisant des développements de Taylor. Les résultats limites sont suffisants pour l'étude asymptotique de l'ACP (Romain, 1979, Dauxois, Pousse et Romain, 1982) ; ce n'est pas le cas pour l'étude asymptotique de l'AC qui nécessite un développement à l'ordre 1 comme nous le montrons ci-après.

5. Applications en Analyse Canonique

Soit $Z = (X, Y)$ un couple de variables aléatoires, d'ordre 4, à valeurs dans $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, espace produit de deux espaces euclidiens $\mathcal{X}$ et $\mathcal{Y}$ munis de la tribu de leurs boréliens. On suppose, sans perte de généralité, que les des v.a.e. X et Y sont centrées et d'opérateurs de covariance l'identité de $\mathcal{X}$ et de $\mathcal{Y}$ respectivement. On pose : $Z = (X, Y)$ et

$$V_Z = \mathbb{E}(Z \otimes Z) = \begin{pmatrix} I_{\mathcal{X}} & V_{XY} \\ V_{YX} & I_{\mathcal{Y}} \end{pmatrix},$$

l'opérateur de covariance de Z. Une définition succincte de l'Analyse Canonique de (X,Y) est la recherche du couple de v.a.e. centrées réduites (f_1, g_1) combinaisons linéaires "normées" des composantes de X et Y respectivement ($f_1 = \langle x_1, X \rangle_{\mathcal{X}}$ avec $\|x_1\|_{\mathcal{X}} = 1$, $g_1 = \langle y_1, Y \rangle_{\mathcal{Y}}$ avec $\|y_1\|_{\mathcal{Y}} = 1$) de corrélation ρ_1 maximale, puis le couple (f_2, g_2), non corrélées aux précédentes, de corrélation ρ_2 maximale, et ainsi de suite.

Les coefficients de corrélation (ρ_i) sont appelés *coefficients canoniques*, les variables (f_i, g_i) *variables canoniques*, les vecteurs (x_i, y_i) *facteurs canoniques.* Les différents éléments de l'Analyse Canonique de (X, Y), appelée AC théorique, sont alors obtenus à partir de la diagonalisation de $R_X = V_{XY} V_{YX}$. Les valeurs propres sont les carrés des coefficients canoniques et les vecteurs propres associés

sont les facteurs canoniques associés. On considère une suite $(Z_i)_{i\in N^\mu}$ de v.a.e. i.i.d. comme Z et on note $\overline{Z}^n$ et

$$V_Z^n = \begin{pmatrix} V_X^n & V_{XY}^n \\ V_{YX}^n & V_Y^n \end{pmatrix}$$

la moyenne d'échantillonnage et l'opérateur de covariance d'échantillonnage. On a alors les théorèmes limites dans $\sigma_2(\mathcal{Z})$: *la suite (V_Z^n) converge p.s. vers V_Z (loi forte des grands nombres) et la suite de terme général $W_Z^n = \sqrt{n}(V_Z^n - V_Z)$ converge en loi vers W_Z, gaussienne centrée d'opérateur de covariance celui de la v.a. $Z \otimes Z$ (théorème de la limite centrale).* Les différents éléments de l'AC d'échantillonnage sont obtenus à partir de la diagonalisation de:

$$R_X^n = (V_X^n)^{-\frac{1}{2}} V_{XY}^n (V_Y^n)^{-1} V_{YX}^n (V_X^n)^{-\frac{1}{2}}.$$

Proposition Dans $\sigma_2(\mathcal{X})$, la suite (R_X^n) converge p.s. vers R_X et la suite de terme général $U_X^n = \sqrt{n}(R_X^n - R_X)$ converge en loi vers la v.a. gaussienne centrée

$$U_X = -\frac{1}{2}(W_X R_X + R_X W_X) + W_{XY} V_{YX} + V_{XY} W_{YX} - V_{XY} W_Y V_{YX}$$

Preuve La preuve utilise le théorème de Rubin-Billingsley (1968) ; il s'agit de trouver une suite d'opérateurs aléatoires (Ψ_X^n) de $\sigma_2(\sigma_2(\mathcal{Z}), \sigma_2(\mathcal{X}))$, telle que l'on puisse écrire : $U_X^n = \Psi_X^n(W_Z^n)$, suite (Ψ_X^n) convergeant p.s. vers l'opérateur non aléatoire Ψ_X de $\sigma_2(\sigma_2(\mathcal{Z}), \sigma_2(\mathcal{X}))$, défini par:

$$\begin{aligned} \Psi_X(T) &= -\frac{1}{2}(P_{\mathcal{X}} T P_{\mathcal{X}}^* R_X + R_X P_{\mathcal{X}} T P_{\mathcal{X}}^*) + P_{\mathcal{X}} T P_{\mathcal{Y}}^* V_{YX} \\ &+ V_{XY} P_{\mathcal{Y}} T P_{\mathcal{X}}^* - V_{XY} P_{\mathcal{Y}} T P_{\mathcal{Y}}^* V_{YX}, \end{aligned}$$

où $P_{\mathcal{X}}$ (resp. $P_{\mathcal{Y}}$) désigne la projection de $\sigma_2(\mathcal{Z})$ sur $\sigma_2(\mathcal{X})$ (resp. $\sigma_2(\mathcal{Y})$). La suite (W_Z^n) convergeant en loi vers W_Z, on en déduit que la suite (U_X^n) converge en loi vers $\Psi_X(W_Z) = U_X$.□

On peut en déduire l'écriture de l'opérateur de covariance de U_X en fonction de celui de W_Z et l'expliciter dans les cas elliptiques ou gaussiens. C'est directement des résultats de la théorie des perturbations que l'on déduit le comportement asymptotique de la suite (u_i^n) de vecteurs propres de U_X^n associée à une suite de valeurs propres convergeant vers une valeur propre simple de U_X dont on note x_i un vecteur propre unitaire associé : la suite $(\sqrt{n}(u_i^n - x_i))$ converge en loi vers $-S_{X_i} U_X x_i$. En revanche, un développement à l'ordre 1 de $\sqrt{n}(u_i^n - x_i)$ est utile pour obtenir le comportement asymptotique du facteur canonique associé x_i^n. Il s'écrit en effet : $x_i^n = (V_X^n)^{\frac{1}{2}} u_i^n$ car il s'agit en fait d'un vecteur propre de l'opérateur:

$$(V_X^n)^{-1} V_{XY}^n (V_Y^n)^{-1} V_{YX}^n (V_X^n).$$

C'est pour se ramener à la diagonalisation d'un opérateur symétrique que nous avons diagonalisé R_X^n.

Proposition Dans $\mathcal{X}$, la suite (x_i^n) converge p.s. vers x_i et la suite $(\sqrt{n}(x_i^n - x_i))$ converge en loi vers la v.a. gaussienne centrée $-\frac{1}{2}W_X x_i - S_{Xi}U_X x_i$.

Preuve On a : $\sqrt{n}(x_i^n - x_i) = \sqrt{n}((V_X^n)^{\frac{1}{2}}u_i^n - x_i) = -(V_X^n)^{-\frac{1}{2}}((V_X^n)^{\frac{1}{2}} + I_{\mathcal{X}})^{-1}\sqrt{n}(V_X^n - I_{\mathcal{X}})u_i^n + \sqrt{n}(u_i^n - x_i)$. Chacun des deux termes converge en loi vers une gaussienne non dégénérée. On ne peut donc conclure qu'en écrivant la loi conjointe en fonction de W_Z^n. On a : $\sqrt{n}(u_i^n - x_i) = -S_{Xi}U_X^n x_i + r_n$
avec (r_n) suite de vecteurs aléatoires de $\mathcal{X}$ convergeant p.s. vers 0 et
$U_X^n = \Psi_X^n(W_Z^n)$. On peut donc trouver une suite (φ_n) d'opérateurs aléatoires à valeurs dans $\sigma_2(\sigma_2(\mathcal{Z}), \mathcal{X})$ telle que l'on puisse écrire:

$$\sqrt{n}(x_i^n - x_i) = \varphi_n(W_Z^n) + r_n,$$

suite (φ_n) convergeant p.s. vers l'opérateur φ défini par:

$$\varphi(T) = -\frac{1}{2}P_{\mathcal{X}}TP_{\mathcal{X}}^* - S_{Xi}\Psi_X(T)x_i.$$

Le théorème de Rubin-Billingsley permet alors de conclure que $(\sqrt{n}(x_i^n - x_i))$ converge en loi vers : $\varphi(W_Z) = -\frac{1}{2}W_X x_i - S_{Xi}U_X x_i$.□

Références

Anderson, T.W. (1963). Asymptotic Theory for Principal Component Analysis. *Ann. Math. Statist.* 34, 122-148.

Anderson, T.W. (1999). Asymptotic study for Canonical Correlation Analysis. *J. Multivariate. Anal,* 70, 1-29.

Bhattacharya, R. et Ghosh J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.*, 6, 434-451.

Dauxois, J. , Pousse, A. et Romain, Y. (1982). Asymptotic theory for the Principal Component Analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12, 136-154.

Fine, J. (1981). *Analyses en Composantes Principales Réduites d'une fonction aléatoire hilbertienne à l'aide de la théorie des perturbations.* Thèse de 3ème cycle, UPS, Toulouse.

Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. *Statistics*, 18, 401-414.

Fine J. (2000). Etude asymptotique de l'Analyse Canonique. *Annales de l'ISUP,* 44, 2-3, 50 pages, à paraître.

Kato T. (1966, 1ère éd.). *Perturbation Theory for Linear Operators.* Springer-Verlag, New-York.

Romain, Y. (1979). *Etude asymptotique des approximations par échantillonnage de l' Analyse en Composantes Principales d'une fonction aléatoire. Quelques applications.* Thèse de 3ème cycle, UPS, Toulouse.

Tests d'hypothèse dans le modèle de régression linéaire fonctionnel

Aldo GOIA

L.S.P., Toulouse et Univ. de Turin
goia@cict.fr, goia@econ.unito.it

Exposé du 12 Mars 2001

Résumé

Soit Y une variable aléatorie (v.a.) réelle et X une v.a. definie sur le même espace probabilisé de Y et à valeur dans l'espace d'Hilbert H de fonctions de carré integrable définies sur un ensemble $\mathcal{T} \subset \mathbb{R}$ que l'on prend borné.

On suppose que l'on peut prédire Y par X selon la relation (Ramsay et Silverman, 1997) :

$$Y = \int_{\mathcal{T}} X(t)\varphi(t)dt + \varepsilon, \tag{15}$$

avec $\varphi \in H$ et ε une v.a. réelle centrée et indépendente de X. Grâce au théoreme de Riesz le modèle (15), peut s'écrire (Cardot et al., 1999) :

$$Y = \Psi(X) + \varepsilon, \tag{16}$$

où Ψ est une forme linéaire continue sur H. Des estimateurs ont été proposés par Cardot et al. (1999a et 1999b) pour le modèle (16) et par Ramsay et Silverman (1997) dans le cadre du modèle (15).

Dans cet exposé on aborde le problème de test pour des hypothèses de type

$$\mathcal{H}_0 : \ \varphi = \varphi_0,$$
$$\mathcal{H}_1 : \ \varphi \neq \varphi_0,$$

où φ_0 est une fonction definie sur $\mathcal{T}$. On considére deux types d'hypothèse nulle:

- $\mathcal{H}_0 : \ \varphi_0(t) = 0, \ \forall t \in \mathcal{T}$ (test de nullité du coefficient fonctionnel)

- $\mathcal{H}_0 : \varphi_0(t) = 0, \forall t \in \mathcal{T}^\star$, où $\mathcal{T}^\star$ est un sous intervalle de $\mathcal{T}$ (test de nullité sur un sous intervalle de $\mathcal{T}$).

D'abord on propose des tests basé sur la statistique du rapport de vraisemblance F sous l'hypothèse de normalité de l'erreur. Puis, sous cette même condition, on donne une approximation de la probabilité de refuser $\mathcal{H}_0$ quand elle est vraie, par la technique proposée par Azzalini et Bowman (1993 et 1997) qui utilise des résultats de Johonson et Kotz (1972) sur les distributions des formes quadratiques pour des variables gaussiennes.

Sous de conditions plus generales (Scheffé 1959) on peut simuler par des techniques de Monte Carlo la loi de la statistique F sous $\mathcal{H}_0$ en permutant les observations (Azzalini et Bowman 1997, Good 1993, Raz 1990) et on estime la probabilité d'erreur du premier type par par la proportion des F simulées que sont plus grandes de la $\hat{F}$ empirique calculée sur les données originales.

Références

Azzalini A., Bowman A. (1997). *Applied Smoothing Techniques for Data Analysis.* John Wiley and Sons, New York.

Azzalini A., Bowman A. (1993). "On the use of nonparametric regression for checking linear relationships". *Journal of Royal Statistical Society*, B, 55, 549-557.

Cardot H., Ferraty F., Sarda P. (1999a). "Functional linear model". *Statistics and Probability Letters*, 45, 11-22.

Cardot H., Ferraty F., Sarda P. (1999b). "Spline estimators for the Functional Linear Model: Consistency, Applications and Splus Implementation". *Rapport UBIA Toulouse*, 1999/1.

Ferraty F., Vieu P. (2000). "Functional Nonparametric Model for Scalar Reponse". Preprint.

Good P. (1993). *Permutation tests. A Pratical Guide to Resampling Methods for Testing Hypotheses.* Springer Series in Statistics. Springer.

Johnson N.L., Kotz S. (1972). *Distibutions in Statistics. Continuous Univariate Distributions. Vol II.* Wiley, New York.

Ramsay J.O., Silverman B.W. (1997). *Functional Data Analysis.* Springer.

Raz J. (1990). "Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach". Journal of American Statistical Association, 85, 132-139.

Scheffé H. (1959). *The analysis of variance.* John Wiley and Sons, New York.

Produits (tensoriels et de convolution) de mesures (aleatoires et spectrales)

Alain BOUDOU et Yves ROMAIN

LSP, Toulouse
boudou@cict.fr romain@cict.fr

Exposé du 19 Mars 2001

Résumé

Un sous-titre possible à cette planche pourrait être : "Une étape de plus vers la tensorisation de la Statistique multidimensionnelle". En effet, l'un des objectifs de ce travail a été la recherche de l'intersection de deux domaines récemment étudiés par les deux auteurs (respectivement [1] à [5] et [9] pour le premier et [10] à [12] pour le second), et il aboutit dans un premier temps (voir les travaux [6] à [8]) à la constatation que les notions de produits tensoriels de variables, de processus et d'opérateurs prennent (ou vont prendre...) une place de plus en plus importante en Statistique multidimensionnelle et notamment fonctionnelle.

Ainsi, dans une première partie, on a rappelé les notions tensorielles habituellement utilisées, leurs propriétés et on a mis en exergue la notion pivotale d'ampliation fonctionnelle d'opérateurs (cf. [11]). La "tensorisation" de la Statistique multidimensionnelle classique est illustrée par deux cas d'extension tensorielle de méthodes factorielles (l'ACP et l'Analyse canonique). On réinsiste, au passage, sur le caractère intrinsèquement opératorielle et tensorielle de la Statistique quantique, ce qui motive aussi l'approfondissement des liens avec les disciplines non commutatives.

Dans la seconde partie, on montre que les résultats obtenus sur les produits (tensoriels et de convolution) de mesures (spectrales et aléatoires) (cf. [6] à [8]) fournissent d'autres exemples de tensorisation via l'introduction de la notion d'ampliation de mesures spectrales. Des exemples d'application à des séries stationnaires sont proposés ainsi que des propriétés fondamentales (comme, par exemple, une formule d'intégration de type Fubini).

Références

Azencott, R. et Dacunha-Castelle, D. (1984). *Séries d'observations irrégulières. Modélisation et prévision.* Techniques stochastiques. Masson.

Birman, M. et Solomjak, M. (1996). Tensor product of a finite number of spectral measures is always a spectral measure. *Integr. Equat. Oper. Th.* **24,** *179-187.*

Boudou, A. (1986). Analyses en composantes principales de données aléatoires. *Cahiers du C. E. R. O.,* **28,** 265-281.

Boudou, A. (2000). Produits de mesures et produits de convolution de mesures spectrales. *Publi. Labo. Stat. Proba.* , **14**-00, 1-19. Univ. P. Sabatier, Toulouse.

Boudou, A. et Dauxois, J. (1995). Principal component analysis for stationary random function defined on locally compact abelian group. *J. Mult. Anal.* , **51** *1-16.*.

Boudou, A. et Romain, Y. (2000a). Produits de mesures spectrales et produit tensoriel fonctionnel. *Publi. Labo. Stat. Proba.* , **19**-00, *1-24.* Univ. P. Sabatier, Toulouse.

Boudou, A. et Romain, Y. (2000b). Convolée de mesures aléatoires. *Publi. Labo. Stat. Proba.* , **21**-00, *1-21.* Univ. P. Sabatier, Toulouse.

Boudou, A. et Romain, Y. (2001). Processus hilbertien associé à la convolée de deux mesures aléatoires. *C. R. Acad. Sci. Paris, t.332, série 1, 1-4, 2001. Statistique.*

Dacunha-Castelle, D. et Duflo, M. (1983). *Probabilités et Statistiques. Problèmes à temps mobile.* Masson.

Dauxois, J. , Romain, Y. et Viguier, S. (1994). Tensor products and Statistics. *Lin. Alg. Appl.,* **210** *59-88.*

Romain, Y. (2000a). Etude des somme, différence et produit tensoriels fonctionnels de deux endomorphismes. *Publi. Labo. Stat. Proba.* , **2**-00, *1-65.* Univ. P. Sabatier, Toulouse.

Romain, Y. (2000b). Eléments tensoriels fonctionnels généralisés . *Publi. Labo. Stat. Proba.* , **18**-00, *1-*30. Univ. P. Sabatier, Toulouse.

Analyses factorielles de densités estimées par noyaux gaussiens

Rachid BOUMAZA

Institut National d'Horticulture, Angers
rachid.boumaza@angers.inra.fr

Exposé du 23 Avril 2001

Résumé

On a étudié (Boumaza 1999) l'analyse en composantes principales et l'analyse discriminante de densités gaussiennes multidimensionnelles, en se basant sur l'affinité L^2 de densités. Pour l'analyse discriminante, on y a proposé des règles d'affectation de type géométrique, minimisant une distance, et probabiliste, maximisant une vraisemblance, en s'appuyant sur l'estimation paramétrique des densités. On considèrera l'extension de ces deux analyses au cas de densités quelconques qu'on peut estimer par noyaux gaussiens ; on développera plus particulièrement l'analyse discriminante avec règle d'affectation de type géométrique et discutera le choix du paramètre de lissage. On note $N_x(\mu, \Sigma)$ la densité de la loi de Gauss de paramètres μ et Σ au point x. Soient X un p-vecteur aléatoire de densité f et $(X_1, \ldots, X_{n_f})$ un n_f-échantillon de variable parente X, l'estimateur de f par noyaux gaussiens s'écrit :

$$\hat{f}(x) \;=\; \frac{1}{n_f} \sum_{i=1}^{n_f} N_x(X_i, h^2 S_f) \tag{17}$$

où S_f est la matrice de variance de l'échantillon $n_f^{-1} \sum_i (X_i - \bar{X})(X_i - \bar{X})'$ et h le paramètre de lissage. Cette technique introduite par Fukunaga (Silverman 1986, p.77) offre l'avantage que même dans le cas multidimensionnel on ne recherche qu'une seule fenêtre optimale. Parmi les mesures d'affinité / divergence / distance entre densités introduites dans la littérature (Cf. Adikhari et Joshi 1956, McLachlan 1992, Zografos 1998 pour des synthèses) l'affinité L^2 par sa propriété de linéarité nous paraît particulièrement intéressante. Soient $(X_1, \ldots, X_{n_f})$ et $(Y_1, \ldots, Y_{n_g})$ deux échantillons de densité marginale respective f et g et de matrice de variance S_f et S_g respectivement, on estime l'affinité L^2 de f et g, égale à $\int fg$, par :

$$\int \hat{f}\hat{g} = \frac{1}{n_f n_g} \frac{1}{(2\pi)^{\frac{p}{2}}} \frac{1}{h^p} \frac{1}{|S_f + S_g|^{\frac{1}{2}}} \sum_i \sum_j e^{-\frac{1}{2h^2}(X_i - Y_j)'(S_f + S_g)^{-1}(X_i - Y_j)} \tag{18}$$

où $\hat{f}(x)$ est donnée par l'expression (17) et $\hat{g}(x)$ par l'expression analogue

$$n_g^{-1} \sum_j N_x(Y_j, h^2 S_g).$$

De cette affinité (18) on peut déduire une estimation de la distance entre f et g ou encore de la distance entre les "densités" normalisées $\frac{f}{\|f\|}$ et $\frac{g}{\|g\|}$. On notera que les deux densités sont estimées en utilisant le même paramètre de lissage. Différents critères de sélection du paramètre de lissage ont été développés (Jones and al. 1996) et parmi les plus utilisés celui de l'erreur quadratique intégrée moyenne (MISE). Dans le cas où f est $N(\mu, I_p)$, la densité gaussienne p-dimensionnelle réduite, et

$$\hat{f}(x) = n_f^{-1} \sum_i N_x(X_i, h^2 I_p).$$

On obtient :

$$MISE = \frac{1}{(2\sqrt{\pi})^p} \left(\frac{1}{n_f} (\frac{1}{h^p} - \frac{1}{(1+h^2)^{\frac{p}{2}}}) + 1 + \frac{1}{(1+h^2)^{\frac{p}{2}}} - 2\sqrt{2}^p \frac{1}{(2+h^2)^{\frac{p}{2}}} \right)$$

Cette formule est une extension au cas multidimensionnel de la valeur de MISE obtenu pour $p = 1$ (Fryer 1976). Son minimum est atteint pour h donné dans le tableau (Table 1). L'analyse discriminante fera intervenir plusieurs densités ;

Table 1: Fenêtre optimale en fonction de n_f et p

n_f	p=1	p=2	p=3	p=4	p=5	p=10
5	.902	.940	.974	1.004	1.031	1.130
10	.758	.804	.844	.881	.914	1.037
50	.519	.575	.624	.669	.709	.864

plutôt que de chercher une fenêtre optimale au sens du critère MISE pour chaque densité, on recherchera empiriquement une fenêtre minimisant le taux d'erreur de classement par validation croisée ou sur échantillon test. Cette recherche empirique sera faite selon un processus itératif ; le tableau précédent fournit une valeur initiale pour démarrer ce processus. On dispose de T tableaux $\mathcal{X}_1, \ldots, \mathcal{X}_T$ à p colonnes. Pour tout t de $\mathcal{T} = \{1, \ldots, T\}$, le tableau $\mathcal{X}_t$ est à n_t lignes ; ces lignes seront considérées comme des réalisations indépendantes d'un p-vecteur aléatoire X_t dont la loi absolument continue par rapport à la mesure de Lebesgue a pour densité f_t.
Sur l'ensemble des indices $\mathcal{T}$ est définie une variable qualitative Y à Q modalités engendrant une partition des densités $\{f_t \ : \ t \in \mathcal{T}\}$ en Q classes $\mathcal{G}_1, \ldots, \mathcal{G}_Q$.

Soit un tableau $\mathcal{X}$ à n_f lignes et p colonnes, considéré comme un n_f-échantillon d'un vecteur X de densité f, le problème est d'affecter la densité f à une des classes $\mathcal{G}_1, \ldots, \mathcal{G}_Q$. Chaque classe $\mathcal{G}_k$ est représentée par une densité g_k construite à partir des densités f_t appartenant à cette classe et estimée à partir de l'échantillon constitué en regroupant les échantillons correspondant à ces mêmes densités. On calcule les distances de f aux g_k puis on affecte f à la classe la plus proche. La règle ainsi bâtie appliquée aux densités f_t dont la classe d'appartenance est connue permet de rechercher "manuellement" le paramètre de lissage h optimal au sens du taux d'erreur minimum ; cette recherche manuelle peut être initialisée en retenant la valeur optimale obtenue ci-dessus (Table 1). La démarche adoptée en analyse discriminante pour l'estimation des densités puis le choix du paramètre peut être étendu au cas de l'analyse en composantes principales en cherchant à maximiser un critère, par exemple le pourcentage d'inertie expliquée par le premier axe principal. On terminera l'exposé par des illustrations.

Références

Adikhari, B.P. et Joshi, D.D. (1956). Distance. Discrimination et résumé exhaustif. *Publications de l'Institut de Statistique de l'Université de Paris*, 57–74.

Boumaza, R. (1999a). *Analyses factorielles des distributions marginales de processus.* Thèse de Doctorat, Université J. Fourier, Grenoble, France.

Fryer, M.J. (1976). Some errors associated with the non-parametric estimation of density functions. *J. Inst. Maths. Applics*, 18, 371–380.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *JASA*, 91 (433), 401–407.

Mc Lachlan, G.J. (1992). *Discriminant analysis and statistical pattern recognition.* Wiley, New-York.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Zografos, K. (1998). f-dissimilarity of several distributions in testing statistical hypothesis. *Ann. Inst. Statis. Math.*, 50 (2), 295–310.