3èmes Journées de Statistique Fonctionnelle et Opératorielle Toulouse les 13-14 Juin 2005

Recueil de résumés

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

2

.

Comité de programme

Alain BOUDOU (UPS), Hervé CARDOT (BIA-INRA, Auzeville), Frédéric FERRATY (UPS/UTM), Aldo GOIA (Fac. d'économie, Novara), André MAS (Univ. Montpellier 2), Mikhail NIKOULINE (Univ. Bordeaux2), Mustapha RACHDI (Univ. Pierre Mendès-France, Grenoble), Juan RODRIGUEZ POO (Univ. de Cantabria, Santander), Yves ROMAIN (UPS), Pascal SARDA (UPS/UTM), Philippe VIEU (UPS), Sylvie VIGUIER-PLA (Univ. Perpignan), Abderrahmane YOUSFATE (Univ. Sidi-Bel-Abbes).

Liste des conférenciers

Jorge BARRIENTOS (Univ. Alicante), Rachid BOUMAZA (INH, Angers), Vincent COUALLIER (Univ. Bordeaux 2), Abdelnasser DAHMANI (Univ. Bejaia), Laurent GARDES (Univ. Pierre Mendès-France, Grenoble), Sonia GRICHE (Univ. Pierre Mendès-France, Grenoble), Ali LAKSACI (Univ. Sidi-Bel-Abbes), David NERINI (Univ. Aix-Marseille 2), Sophie NIANG (Univ. Lille 3), Besnik PUMO (INH, Angers), Alejandro QUINTELA DEL RIO (Univ. La Corogne), Yves ROMAIN (UPS), Ernesto SALINELLI (Univ. degli studi del Piemonte Orientale, Novara).

	Troisièmes Journées de Statistique fonctionnelle et opératorielle Toulouse - 13 ***et*** 14 ***juin*** 2005 -		***PROGRAMME***
Horaires	**Lundi** 13 **juin** 2005	Horaires	**Lundi** 13 **juin** 2005
		14h – 14h30	**D. NERINI** (Univ. Marseille) *Arbres de régression fonctionnels. Construction et application en océanologie.*
9h45 – 10h	*Ouverture et Présentation des Journées*	14h30 – 15h	**A. LAKSACI** (Univ. Sidi-bel-Abbès, Algérie et UPS, Toulouse) *Analyse non-paramétrique de données fonctionnelles*
10h – 10h45	**E. SALINELLI** (Fac. d'économie, Novara, Italie) *Nonlinear Principal Component Analysis.*	15h – 15h30	**S. NIANG** (Univ. Lille III) *Estimation non-paramétrique de la régression pour des champs aléatoires à valeurs dans un espace métrique.*
	PAUSE		PAUSE
11h – 11h30	**V. COUALLIER** (Univ. Bordeaux II) *Degradation modelling and semiparametric estimation of survival characteristics of degradation dependent failure times*	16h – 16h30	**B. PUMO** (INH, Angers) *The ARHD model*
11h30 – 12h	**S. GRICHE** (Univ. Pierre Mendès-France, Grenoble) *Functional data: nonparametric estimation of the regression function for dependent error process*	16h30 – 17h	**J. BARRIENTOS-MARÍN** (Univ. Alicante) *Local Linear Weighted Regression for Functional Data*

2

	Université Paul Sabatier ***Amphi Laurent Schwarz*** ***Bâtiment*** $1R3$
Horaires	**Mardi** 14 **juin** 2005
9h30 – 10h15	**Y. ROMAIN** (Univ. P. Sabatier, Toulouse) *Some recent developments in operator-based statistics.*
10h15 – 10h45	**R. BOUMAZA** (INH, Angers) *Analyse en Composantes Principales et analyse discriminante de densités de probabilité dans l'environnement R*
	PAUSE
11h – 11h30	**A. QUINTELA del RIO** (Univ. Santiago, Espagne) *Plug-in bandwidth selection in nonparametric hazard estimation: Seismology applications*
11h30 – 12h00	**L. GARDES** (Univ. P. Mendès-France, Grenoble) *Estimation d'une fonction quantile extrême*
12h00 – 12h30	**A. DAHMANI** (Univ. de Béjaia) *Approche statistique pour la résolution d'un problème mal posé*

6

.

3^{emes} Journées de Statistique Fonctionnelle et Opératorielle Toulouse, 13-14 Juin 2005

Alain BOUDOU, Hervé CARDOT, Frédéric FERRATY
Yves ROMAIN, Pascal SARDA, Philippe VIEU
et Sylvie VIGUIER-PLA

Coordinateurs du groupe de travail STAPH
Laboratoire de Statistique et Probabilités
Toulouse

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

La statistique fonctionnelle et opératorielle occupe désormais une place importante dans la recherche en statistique : les thèmes les plus actuels dans ce domaine ont trait à la modélisation statistique pour variables fonctionnelles. Cet intérêt provient autant du large potentiel d'applications (imagerie, télédétection, météorologie, médecine, ..) que des problèmes théoriques qu'elle engendre. La statistique fonctionnelle et opératorielle connaît donc un essor à l'échelle internationale et c'est particulièrement le cas à Toulouse. Ainsi, des chercheurs membres du LSP et pour certains du département BIA de l'INRA et de l'équipe GRIMM de l'UTM animent sur ce thème le groupe de travail STAPH dont le prolongement a été l'organisation de rencontres en juin 2002 et juin 2003 à l'UPS mais également de sessions lors du premier congrès Canada-France des Sciences mathématiques à Toulouse en juillet 2004, de COMPSTAT 2004 à Prague en août 2004 et du congrès IASC-2005 qui se déroulera en octobre à Chypre.[1]

Les rencontres, dont les résumés des exposés sont présentés dans ce document, sont donc la troisième édition de cette manifestation. Ces journées sont, dans la continuité des précédentes, destinées à promouvoir la statistique fonctionnelle et opératorielle, à travers la rencontre de chercheurs du domaine venant de laboratoires français et étrangers. Par ailleurs, l'accent a été principalement mis cette année sur la participation de jeunes statisticiens et sur une diversité d'approches balayant un champ large de la théorie aux applications.

Ces journées s'inscrivent également dans le cadre de collaborations avec d'au-

[1] Toutes nos activités sont accessibles sur la page web
http : //www.lsp.ups − tlse.fr/staph.html

tres Universités. Il s'agit tout d'abord de l'Université de Sidi-bel-Abbès, Algérie, avec laquelle une collaboration a été initiée il y a cinq ans (deux thèses en cotutelle avec cette Université sont en cours). L'équipe de statisticiens de cette Université a par ailleurs participé aux précédentes Journées (2002 et 2003) qu'elle a co-organisées. D'autres Universités sont également associées à l'organisation de cette troisième édition de nos journées par le biais de la participation de chercheurs au comité de programme. Il s'agit des Universités de Bordeaux 2, Montpellier 2, Grenoble (Pierre Mendès-France), Novarra (Italie) et Santander (Espagne).

Nous souhaitons remercier vivement les différents organismes ayant apporté leur aide financière à cette manifestation : il s'agit de l'Université Paul Sabatier, du Laboratoire de Statistique et Probabilités, du département BIA de l'INRA-Auzeville, de l'équipe GRIMM de l'Université du Mirail et du Conseil Régional de Midi-Pyrénées .

Local Linear Weighted Regression for Functional Data

Jorge BARRIENTOS*, Frédéric FERRATY et Philippe VIEU

* Adresses pour correspondance :
Departamento de Fundamentos de Analisis Economico
Universidad de Alicante
Alicante, Espagne
et
Laboratoire de Statistique et Probabilités
Université Paul. Sabatier
Toulouse, France

e-mail : barrient@cict.fr ou jbarr@merlin.fae.ua.es

Abstract

The aim of this paper is to extend the functional nonparametric methods to local linear weighted regression with scalar response. In this stage, the Nadaraya-Watson is a local constant weighted estimator and appear like a particular case of our linear weighted fits. In this double infinite framework, the explanatory variable is valued in some abstract semi-metric functional space. Asymptotic behaviour of the estimator will be studied by mean of almost complete convergence results. These one are similar to the standard Nadaraya-Watson.

References

– Cardot, H. F. Ferraty and P. Sarda (1999) Functional Linear Model. *Statistics and Probability Letters*, **45**, 11-22.
– Fan, J (1992). Desing-Adaptive Nonparametric Regression. *Journal of the American Statistical Associations*, **87**, 420, 998-1004.
– Fan, J (1993). Local Linear Regression Smoothers and Their Minimax Efficiences. *Annals of Statistics*, 21, **1**, 196-216.
– Ferraty, F. A. Goia and P. Vieu (2002) Functional Nonparametric Model for Time Series : a Fractal Approach to Dimension Reduction. *TEST*, **11**, 2, 317-344.

- Ferraty, F and P. Vieu (2002) The Functional Nonparametric Model and Application to Spectrometric Data. *Comput. and Statistics*, **17**, 545-564.
- Ferraty, F and P. Vieu (2004). Nonparametric Models For Functional Data, with Applications in Regression, Time Series Prediction and Curve Discrimination. *Non Parametric Statistics*, 16, **1-2**, 111-125.
- Ferraty, F and P. Vieu (2005). Nonparametric Methods for Functional Data. Methods, Theory, Applications and Implementatios. Springer (In print).
- Ramsay, J. O and C. J Dalzell (1991). Some Tools for Functional Data Analysis. *Journal of the Royal Statistics Society, Serie B*, **53**, 3, 539-572.
- Ramsay, J. O and B. W. Silverman (1997). Functional Data Analysis. Springer-Verlang.
- Ruppert, D and M. P. Wand (1994) Multivariate Locally Weigthed Least Squares Regression. *Annals of Statistics*, 22, **3**, 1346-1370.
- Wand, M. P and M. C. Jones (1995) Kernel Smoothing. Monographs on Statistics and Applied Porbability, 60. Chapman & Hall.

Analyse en composantes principales et analyse discriminante de densités de probabilité dans l'environnement R

R. BOUMAZA*, P. GUILERMIN, P. REVOLLON, L. DURANDET

* Adresse pour correspondance :
UMR SAGAH, Institut National d'Horticulture
2 rue Le Nôtre, 49045 Angers, France

e-mail : Rachid.Boumaza@inh.fr

Introduction

On considère des données ternaires " *individus* × *variables* × *occasions* " (Tab. 2) où à chaque occasion t ($t = 1, \ldots, T$), on observe les p mêmes variables quantitatives sur un lot de n_t individus. Ces n_t observations sont considérées comme un n_t-échantillon d'un vecteur aléatoire à valeurs dans R^p, de densité de probabilité f_t (par rapport à la mesure de Lebesgue), qui permet d'estimer cette densité.
L'objectif de l'analyse en composantes principales (ACP) fonctionnelle de densités de probabilité est de visualiser ces occasions, via les densités associées, sur un sous-espace de dimension réduite afin d'apprécier les différences et ressemblances entre lots d'individus. Si t fait référence au temps, cette ACP permettra d'apprécier qualitativement l'évolution des lots d'individus.
Aux données précédentes, on ajoute une variable qualitative G à Q modalités, définie sur l'ensemble des T occasions. Un nouveau lot $T+1$ de n individus sur lesquels on a observé les p variables quantitatives, se présente ; on cherche à prédire la valeur de G pour ce nouveau lot. C'est l'objectif de l'analyse discriminante (AD) de densités ([BOU 04]).

La fonction FPCAd

Présentation L'ACP des densités f_t ($t = 1, \ldots, T$) dans l'espace $L^2(R^p)$ ([BOU 98, KNE 01]) permet d'obtenir la décomposition des f_t suivant un système or-

thonormé (h_k) :

$$f_t = \sum_k \alpha_{kt} h_k \ .$$

Cette décomposition permet ainsi la visualisation de ces densités respectant au mieux les distances entre ces densités ; ces distances sont calculées en s'appuyant sur la norme L^2 induite par le produit scalaire classique de $L^2(R^p)$.
Dans le cas de densités gaussiennes, on peut trouver en [BOU 98] des relations entre cette ACP de densités et la première étape, ou étape de l'interstructure, de la méthode STATIS duale.
La fonction FPCAd de l'environnement R qui réalise cette ACP de fonctions de densités, calcule les coordonnées, les aides à l'interprétation classiques (contributions et qualités) suivant chaque axe principal, et réalise les représentations graphiques.
Les options classiques de l'ACP : centrage ou réduction, peuvent être sélectionnées. Les densités peuvent être considérées soit gaussiennes et estimées paramétriquement, soit quelconques et estimées par la méthode du noyau. Pour la méthode du noyau on utilise le noyau gaussien avec la fenêtre AMISE : $w = (4/(n(p+2)))1/(p+4)$ ([SIL 86]) qui minimise une approximation de l'erreur quadratique moyenne intégrée (MISE).

Description de la fonction. La fonction est définie comme suit :
FPCAd ← function(X, gaussian=T, centered=F, normed=T, nb.factors=3, nb.values=10, save.results=F, filename.results="FPCAd_results.RDATA").

Cette fonction a pour entrées :

- un data frame à $(p+1)$ colonnes :
  - les p premières colonnes sont les variables quantitatives
  - la $(p+1)$-ième colonne est la variable désignant le facteur occasion dont le nom est obligatoirement "Lot".
- une variable logique : gaussian (TRUE, par défaut) si on suppose ou non la normalité ; si FALSE, on utilise le noyau gaussien avec fenêtre AMISE.
- deux variables logiques : centered (TRUE, par défaut) et normed (FALSE, par défaut), pour indiquer si l'ACP doit être centrée et/ou normée.
- un nombre nb.factors (3 par défaut) indiquant le nombre de facteurs principaux retenus. Si le nombre de lots est inférieur à nb.factors, c'est le nombre de lots qui sera utilisé.
- un nombre nb.values (10 par défaut) indiquant le nombre de valeurs principales à afficher. Si le nombre de lots est inférieur à nb.values, c'est le nombre de lots qui sera utilisé.
- une variable logique : save.results (FALSE par défaut) qui demande la sauvegarde de tous les résultats (inerties expliquées, coordonnées, qualités et contributions) dans un fichier dont le nom peut-être choisi par l'option suivante.

- une variable caractère : filename.results qui indique le nom du fichier (de type .RDATA) où seront sauvegardés les résultats (inerties, coordonnées, qualités et contributions) ; par défaut le nom du fichier est "FPCAd_results.RDATA".

Cette fonction a par défaut les sorties suivantes :

- les graphiques sur les premiers plans principaux (par défaut les 3 premiers) ; une fenêtre graphique (device) par plan principal, les 3 fenêtres sont superposées.
- la liste, désignée ci-après par le symbole ..., comprenant :
  - ...[[1]] : les inerties expliquées,
  - ...[[2]] : les coordonnées des lots sur les nb.factors (3 par défaut) premiers axes,
  - ...[[3]] : les qualités carrées correspondantes par axe
  - ...[[4]] : et enfin les contributions correspondantes.

La fonction FDAd

Présentation. L'AD de densités de probabilité permet d'affecter le lot $T+1$ à l'une des modalités de la variable G. Dans [BOU 04], il a été proposé des règles géométriques et des règles probabilistes. Comme en ACP de densités, à chaque lot t est associée une densité f_t ; de plus à chaque modalité q de G est associée une densité g_q qu'on estime à partir des lots qui prennent cette modalité. Les règles géométriques affectent le lot $T+1$ de densité f à la modalité la plus proche au sens d'une mesure de dissimilarité. Les règles probabilistes quant à elles, supposent que les densités f_t et g_q sont gaussiennes et calculent une probabilité d'affectation du lot à chaque modalité, l'affectation pouvant alors se faire à la modalité pour laquelle la probabilité est maximum. La fonction FDAd de l'environnement R qui réalise les calculs des mesures de dissimilarité ou probabilités comporte plusieurs options. Elle est paramétrée selon le type de données disponibles et la règle d'affectation à utiliser :

- Le type de données :
  - Le nombre T_q de lots par modalité q : $T_q > 1$ $(\forall q)$ ou $T_q = 1$ $(\forall q)$.
  - Les densités $(f_t)_{t=1,T}$ et $(g_q)_{q=1,Q}$ respectivement associées aux T lots et aux Q modalités sont considérées gaussiennes ou non.
- Le choix de la règle d'affectation tient compte du type de données et doit préciser :
  - La méthode d'estimation des densités : méthode paramétrique ou méthode du noyau (non paramétrique), et pour cette dernière méthode le type de fenêtre de lissage : fenêtre AMISE par densité ou fenêtre commune à toutes les densités.
  - La mesure de dissimilarité entre densités : distance L^2, distance L^2 après normalisation des densités, distance de Matusita ou mesure de Jeffreys

(divergence de Kullback-Leibler symétrisée).
– Le critère utilisé : géométrique ou probabiliste.

Certains choix parmi les options décrites sont bien entendu incompatibles. Le tableau 1 présente les choix possibles.
Dans le cas où on dispose de plusieurs lots par modalité ($T_q > 1,\ \forall q = 1, \ldots, Q$), les différents règles d'affectation peuvent être comparées sur la base des taux de bon classement obtenus par validation croisée. Ce qui permet de discuter le choix d'une "meilleure" option pour l'affectation du lot $T + 1$.

Description de la fonction. La fonction est définie comme suit :
FDAd ← function(G, X, lots.per.group=1, gaussian=T, window=NULL, dissimilarity.measure="L2", save.results=F, filename.results="FDAd_results.RDATA")

Cette fonction a pour entrées :

- un premier data frame à 2 colonnes :
  - la première nom.Lot contient la liste des lots
  - la deuxième colonne qualite.Lot donne pour chaque lot sa qualité.
- un deuxième data frame à (p+1) colonnes :
  - les p premières colonnes sont les variables quantitatives
  - la (p+1)-ième colonne est la variable désignant le facteur "Lot", c'est-à-dire le nom du lot auquel appartient la ligne.
- une variable numérique : lots.per.group pouvant prendre les valeurs 1 (si à une modalité de la variable qualité on associe une seule densité, qu'on estime à partir d'un seul échantillon) ou 2 (si à une modalité de la variable qualité on associe plusieurs densités qui sont moyennées). Par défaut : lots.per.group=1.
- une variable logique : gaussian si on suppose ou non la normalité ; si FALSE, on utilise le noyau gaussien. Par défaut : gaussian=TRUE
- une variable numérique : window qu'on utilise si gaussian vaut FALSE. C'est la valeur de la fenêtre de lissage qui est utilisée pour toutes les densités. Si la valeur de window est mise à NULL, alors on utilise une fenêtre AMISE par densité. Par défaut : window=NULL.
- une variable caractère : dissimilarity.measure qui fixe la méthode de calcul des dissimilarités entre densités. Elle peut prendre les valeurs suivantes : "L2", "L2N", "M" (ou "Matusita"), "J" (ou "Jeffreys"). Par défaut : dissimilarity.measure="L2".
- une variable logique : save.results (FALSE par défaut) qui demande la sauvegarde de tous les résultats (inerties expliquées, coordonnées, qualités et contributions) dans un fichier dont le nom peut-être choisi par l'option suivante.
- une variable caractère : filename.results qui indique le nom du fichier où seront sauvegardés les résultats) ; par défaut le nom du fichier est "FPCAd_results.RDATA".

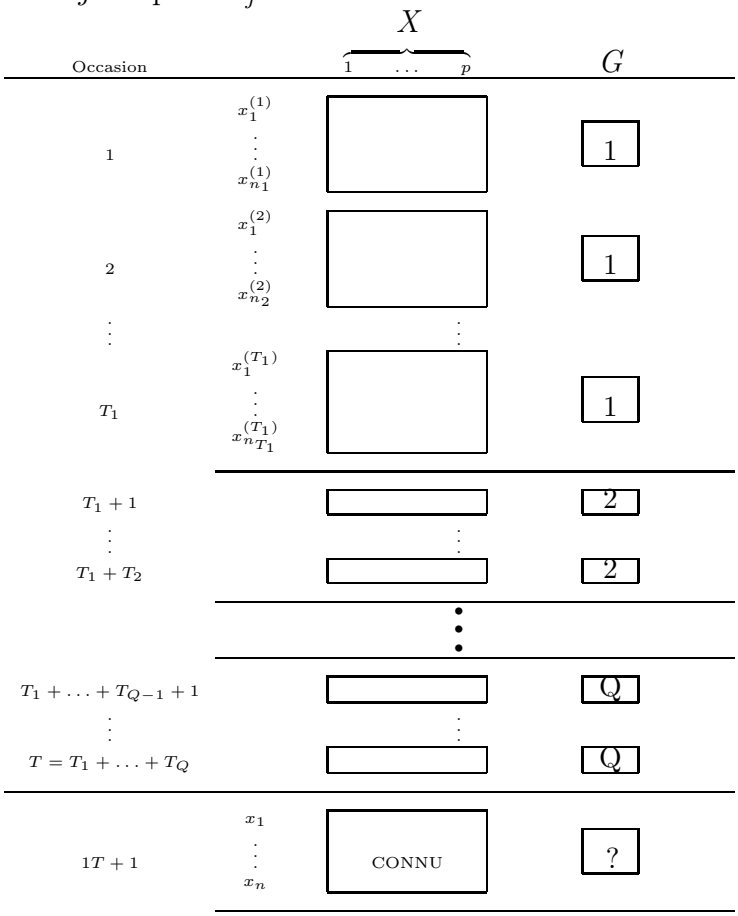
Cette fonction a pour sorties la liste, désignée ci-après par le symbole ..., comprenant :

- ...[[1]] : les dissimilarités entre lots de qualité inconnue et modalités de la variable qualité du lot,
- ...[[2]] : les dissimilarités en pourcentages,
- ...[[3]] : la modalité réalisant le minimum de ces dissimilarités.

TAB. 1 – Compatibilité des choix possibles en analyse discriminate de densités.

<table>
<tr><th colspan="2">Type de données</th><th colspan="3">Règle d'affectation</th></tr>
<tr><th>Nombre d'occasions</th><th>Hypothèse de normalité</th><th colspan="2">Méthode d'estimation</th><th>Mesure de dissimilarité</th><th>Critère</th></tr>
<tr><td rowspan="2">$\forall j,\ T_j > 1$</td><td>Non</td><td>Noyau</td><td>Fenêtre optimale AMISE</td><td rowspan="2">L^2 ou sa version normalisée</td><td rowspan="2">- Géométrique
- g_q est la moyenne des densités f_t prenant la modalité q</td></tr>
<tr><td>Oui</td><td colspan="2">Paramétrique</td></tr>
<tr><td rowspan="4">$\forall j,\ T_j = 1$</td><td rowspan="2">Non</td><td rowspan="2">Noyau</td><td rowspan="2">AMISE</td><td>L^2 ou sa version normalisée</td><td>Géométrique</td></tr>
<tr><td rowspan="3">Kullback-Leibler
Matusita
L^2 normalisée</td><td rowspan="3"></td></tr>
<tr><td rowspan="3">Oui</td><td colspan="2" rowspan="3">Paramétrique</td></tr>
<tr><td>L^2</td><td>Géométrique ou probabiliste</td></tr>
<tr><td colspan="2">Règle quadratique généralisée</td></tr>
</table>

TAB. 2 – Données ternaires "*individus* × *variables* × *occasions*". A la première occasion, les n_1 observations correspondent à un n_1 échantillon de X, à la seconde occasion les n_2 observations correspondent à un n_2-échantillon de X, etc. A chaque occasion, la variable G prend ses valeurs dans $\{1, \ldots, Q\}$, et pour chaque $j = 1, \ldots, Q$, la valeur j est prise T_j fois.



Références

[BOU 98] BOUMAZA, R. (1998) Analyse en composantes principales de distributions gaussiennes multidimensionnelles. *Revue de Statistique Appliquée*, vol. XLVI, n° 2, p. 5-20.

[BOU 04] BOUMAZA, R. (2004) Discriminant analysis with independently repeated multivariate measurements : an L^2 approach. *Computational Statistics & Data Analysis*, vol. 47, p. 823-843.

[KNE 01] KNEIP, A., UTIKAL, K.J. (2001) Inference for density families using functional principal component analysis. *Journal of the American Statistical Society*, vol. 96, n° 954, p. 519-542.

[R 04] R DEVELOPMENT CORE TEAM (2004) *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL : http ://www.R-Project.org.

[SIL 86] SILVERMAN, B.W. (1986) *Density estimation for statistics and data analysis*. Chapman & Hall, London.

18

.

Degradation modeling and semiparametric estimation of survival characteristics of degradation-dependent failure times

Vincent COUALLIER *

* Adresse pour correspondance :
Equipe Statistique Mathématique et ses Applications EA2961
U.F.R. Sciences et Modélisation
Université Victor Segalen Bordeaux 2
Bordeaux

e-mail : couallier@sm.u-bordeaux2.fr

Résumé

In survival analysis, regression models and conditional definitions of the hazard function are used to take into account the effects of the environment and of individual frailties, possibly due to degradation of the item. These effects are modelled by explanatory covariates with increasing complexity such that dummy variables, real valued random covariates or stochastic processes. For instance, conditionally on the random vector A, the famous Proportional Hazard rate model by Cox specifies that the hazard rate of a life time T verifies $\lambda_T(t|A) = \lambda_o(t) \times e^{\beta^T A}$ where λ_o is a unknown baseline hazard function and β is a vector of parameters. This model was first defined for constant in time covariate but it is possible to allow a time-varying effect of the environment on the survival of the item. Wulfsohn and Tsiatis (1997) study the model $\lambda_T(t|A) = \lambda_o(t) \times e^{\beta(A_1 + A_2 t)}$ where $A = (A_1, A_2)$ is some random but fixed in time vector of coefficients. Also, this unit-to-unit variability in the definition of the hazard rate can be interpreted as an individual frailty, modelled by a stochastic process and reflecting an internal accumulation of wear called aging or degradation process.

Bagdonavicius and Nikulin (2004) define the conditional hazard rate given the random vector A as $\lambda(t|A) = \lambda_o(t) \times \lambda(g(t, A))$ where g is a given non decreasing function. Is is strongly related to the model of Wulfsohn and Tsiatis but the assumptions made for estimation and inference are completely different.

Let us assume now that the degradation of an item is given by the sample path of a non decreasing real-valued right continuous and left hand limited stochastic

process $Z(t)$, $t \in I$. Lawless and Crowder (2004) and Couallier (2004) consider gamma processes, Kahle and Wendt (2004) consider marked point processes and Doksum and Normand (1995), Whitmore (1995) and Whitmore and Schenkelberg (1997) consider gaussian processes. As in Lu and Meeker (1993), Meeker and Escobar (1998) and Bagdonavicius *et al* (2004), we make here the assumption that the unknown degradation process is

$$Z(t) = g(t, A), t > 0, \tag{1}$$

where g is a differentiable and non decreasing parametric function of the time and A is a random variable in $\mathbb{R}^p$ which takes account on the variability of the degradation evolution. The degradation values can be absolutely unknown (such that a latent variable) or partially measured, often with error measurements.

This degradation process leads to two possible modes of failure. The first one consists in failure time defined as hitting time of the degradation process i.e. the life time T_0 is the first time of crossing a ultimate threshold z_0 (which can be random) for Z(t)

$$T_0 = \inf\{t \in I, Z(t) \geq z_0\}.$$

The failure time T_0 is sometimes called soft failure (or failure directly due to wear) because in most of industrial applications, z_0 is fixed and the experiment is voluntarily ceased at the time the degradation process reaches the level z_0 or just after this time.

The second one considers that the degradation process influences the distribution of a traumatic failure time T through a conditional definition of its survival function.

$$P(T > t|Z(s), 0 \leq s \leq t) = \exp\left(-\int_0^t \lambda_T(Z(s)))ds\right). \tag{2}$$

In that case, the "hazard rate" of the failure time T does not live in the "time domain" but in the degradation domain. This is very intuitive because an item with low degradation will be at low risk and vice versa.

When both modes of failure exist, only the first one is observed. Very often, longitudinal observations of degradation values (measured with error) are available for each item until the first failure. We are interested here in the non-parametric estimation of the cumulative intensity function $\Lambda(z) = \int_0^z \lambda_T(u)du$ of the traumatic failure T.

In this presentation, the degradation process is modelled by a nonlinear mixed effect regression model. Three different assumptions for the noises are considered (i.i.d., Wiener process, Continuous autoregressive process). The regression parameters which are random from unit to unit are distribution free and are estimated by a generalized nonlinear least squares minimization. A Nelson-Aalen-type estimate of the cumulative intensity function of T is studied. The choices of the

numerical procedure in the statistical estimation and of the covariance structure of the noise are investigated.

Références

Bagdonavicius V, Bikelis A, Kazakevicius V. (2004). Statistical analysis of linear degradation and failure time data with multiple failure modes. Lifetime Data Anal. **10**(1), 65-81.

Bagdonavicius, V., Nikulin, M. (2004). Semiparametric analysis of degradation and failure time data with covariates, in *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* Series : Statistics for Industry and Technology Nikulin, M.S. ; Balakrishnan, N. ; Mesbah, M. ; Limnios, N. (Eds.), Birkauser.

Couallier, V. (2004). Comparison of parametric and semiparametric estimates in a degradation model with covariates and traumatic censoring in *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* Series : Statistics for Industry and Technology Nikulin, M.S. ; Balakrishnan, N. ; Mesbah, M. ; Limnios, N. (Eds.), Birkauser.

Doksum K.A., Normand S.L. (1995). Gaussian models for degradation processes-Part I : Methods for the analysis of biomarker data. Lifetime Data Anal., **1** (2), 131-44.

Kahle, W., Wendt H. (2004). On a cumulative damage process and resulting first passages times, Applied Stochastic Models in Business and Industry **20**(1) : 17-26.

Lawless J, Crowder M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure, Lifetime Data Anal. 10(3) : 213-27.

Lu, C.J. , and Meeker, W.Q. (1993). Using degradation measures to estimate a time-to-failure distribution, Technometrics, **35**, 161-174.

Meeker, W.Q. and Escobar, L. (1998). *Statistical Analysis for Reliability Data*, John Wiley and Sons, New York.

Whitmore GA. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error, Lifetime Data Anal. ; **1**(3) :307-19.

Whitmore GA, Schenkelberg F. (1997). Modelling accelerated degradation data using Wiener diffusion with a time scale transformation, Lifetime Data Anal. ; **3**(1) :27-45.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error, Biometrika, **53**, 330-339.

Approche statistique pour la résolution d'un problème mal posé

Abdelnasser DAHMANI* et Ahmed AIT SAIDI

* Adresse pour correspondance :
Laboratoire de Mathématiques Appliquées
Université de Béjaia 06000
ALGERIE

e-mail : a-dahmani@yahoo.fr

Résumé

Divers domaines des sciences appliquées posent des problèmes dont la résolution exige un certain formalisme mathématique, ce qui a poussé quelques mathématiciens de renom comme Hadamard puis Tikhonov à définir les conditions que doit vérifier chaque problème pour qu'il soit correctement posé. Les problèmes ne rentrant pas dans ce cadre sont appelés "**problèmes mal posés**".

Dans ce travail, nous considérons un cadre classique de problèmes mal posés qui consiste en l'équation à opérateur $Ax = u$ où A est un opérateur défini sur un espace de Hilbert, à inverse non continu. A la différence des méthodes déterministes, nous supposons que le second membre est observé avec des erreurs aléatoires. Nous proposons des méthodes itératives pour la résolution de ce type de problèmes.

.

Estimation d'une fonction quantile extrême

Laurent GARDES

* Adresses pour correspondance :
Université Grenoble 2,
LabSAD, 1251 avenue centrale, B.P. 47
38040 Grenoble Cedex 9, France

e-mail : Laurent.Gardes@upmf-grenoble.fr

Introduction

Soit (X, Y) un couple de variables aléatoires de support :

$$S = \{(x, y) \in \mathbb{R} | 0 \leq x \leq 1 \text{ et } 0 \leq y \leq g(x)\},$$

où g est une fonction inconnue que l'on suppose continue. Disposant de n couples de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendants et de même loi que le couple (X, Y), on désire estimer la **fonction quantile extrême** :

$$\begin{array}{rl} g_{p_n} : & [0,1] \to [0,\infty) \\ & x_0 \mapsto g_{p_n}(x_0) \end{array}$$

où $g_{p_n}(x_0)$ est défini par :

$$\mathrm{P}[Y \leq g_{p_n}(x_0) | X = x_0] = 1 - p_n,\ p_n < 1/n.$$

Le fait que p_n soit strictement inférieur à $1/n$ implique qu'avec une probabilité qui tend vers 1 lorsque n tend vers l'infini, toutes les observations sont situées au-dessous de la fonction g_{p_n}. Il ne s'agit donc pas ici d'un problème d'estimation de quantile conditionnel classique puisqu'il va falloir extrapoler au-delà de l'observation maximale.
L'estimation d'une fonction quantile extrême trouve de nombreuses applications notamment en hydrologie (détermination des limites d'une zone inondable), en économétrie, en analyse d'images, etc ...Dans la littérature, on trouve essentiellement des estimateurs de la frontière g du support (cas particulier où $p_n = 0$). Citons entre autres Gijbels et Peng (2000) qui proposent la méthode suivante

pour estimer $g(x_0)$. Ils utilisent les observations dont la première coordonnée est "suffisamment" proche de x_0. Plus précisément, ils considèrent les variables aléatoires :

$$\{Z_i = Y_i \mathbb{1}\{|X_i - x_0| \leq h_n\},\ i = 1, \ldots, n\},$$

où (h_n) est une suite strictement positive qui converge vers zéro lorsque n tend vers l'infini. Dans la suite, on note $Z_{1,n} \leq \ldots \leq Z_{n,n}$ les variables aléatoires ordonnées. En supposant que la densité du couple (X, Y) est de la forme :

$$f(x,y) = a(x,y)\{g(x) - y\}^{-1/\xi(x)-1} \left\{1 + O\left((g(x) - y)^{\beta(x)}\right)\right\}, \tag{3}$$

où $a(x, y) > 0$, $\xi(x) < 0$ et $\beta(x) > 0$, ils proposent d'estimer $g(x_0)$ soit par

$$\hat{g}_n(x_0) = Z_{n,n},$$

soit, en utilisant l'estimateur du point terminal proposé dans un cadre unidimensionnel par Dekkers et de Haan (1989), par :

$$\tilde{g}_n(x_0) = \frac{Z_{n-k,n} - Z_{n-2k,n}}{2^{-\hat{\xi}_n(x_0)} - 1} + Z_{n-k,n},$$

où $k < n/4$, $k \to \infty$ et $k/n \to 0$ et

$$\hat{\xi}_n(x_0) = \log\left(\frac{Z_{n-k,n} - Z_{n-2k,n}}{Z_{n-2k,n} - Z_{n-4k,n}}\right) / \log(2).$$

Definition de l'estimateur

En utilisant aussi les variables aléatoires $\{Z_i,\ i = 1, \ldots, n\}$ et en supposant que la fonction de répartition conditionnelle de Y sachant que $X = x$ est définie par :

$$F_Y^x(y) = \mathrm{P}[Y \leq y | X = x] = 1 - (g(x) - y)^{-1/\xi(x)}\, \ell_x\left((g(x) - y)^{-1}\right), \tag{4}$$

on propose ici d'estimer $g_{p_n}(x_0)$ par

$$\check{g}_{p_n}(x_0) = Z_{n,n} \frac{v_n\{\tau_{u_n}^{-\hat{\xi}_n(x_0)} - (np_n)^{-\hat{\xi}_n(x_0)}\} - u_n\{\tau_{v_n}^{-\hat{\xi}_n(x_0)} - (np_n)^{-\hat{\xi}_n(x_0)}\}}{\tau_{u_n}^{-\hat{\xi}_n(x_0)} - \tau_{v_n}^{-\hat{\xi}_n(x_0)}},$$

où (u_n) et (v_n) sont deux suites de $]0, 1[$ qui convergent vers 1 lorsque n tend vers l'infini,

$$\tau_{u_n} = \sum_{i=1}^{n} \mathbb{1}\{Z_i \geq u_n Z_{n,n}\},$$

et $\hat{\xi}_n(x_0)$ est un estimateur faiblement consistant de $\xi(x_0)$. Sous des conditions de régularité sur la fonction de répartition conditionnelle F_Y^x et sur la fonction g, on démontre que pour tout $x_0 \in]0,1[$,

$$g_{p_n}(x_0) - \check{g}_{p_n}(x_0) \to_P 0.$$

Des simulations permettent de montrer que l'estimateur $\check{g}_{p_n}$ donne de meilleur résultats que les estimateurs $\hat{g}_{p_n}$ et $\tilde{g}_{p_n}$ proposés par Gijbels et Peng. A noter aussi que l'on peut montrer que la condition (4) sur la loi du couple (X, Y) est plus faible que la condition (3).

References

– Gijbels, I. and Peng, L. (2000). Estimation of a support curve via order statistics, *Extremes*, **3**, 251-277.

– Dekkers, A.L.M. and de Haan, L. (1989). On the estimation of the extreme value index and large quantile estimation, *Annals of Statistics*, **17**, 1795-1832.

28

.

Functional data : Nonparametric estimation of the regression function for dependent error process

Sonia HEDLI-GRICHE

* Adresse pour correspondance :
Université de Grenoble
UFR SHS, BP. 47
F38040 Grenoble

e-mail : Sonia.Griche@upmf-grenoble.fr

Abstract

In this paper we study the nonparametric regression model with stationary dependent errors, and when the explanatory variable is of functional type. We give the optimal convergence rates of the nonparametric kernel estimator of the regression function in the mean square and almost sure sense under general conditions. In particular, we derive similar asymptotic results for long range dependent error process. Some simulation studies are generated when the error process is a fractional brownian motion.

Functional nonparametric regression

The main aim of this paper is to study the estimation of the regression operator r in the following model :

$$Y_i \;=\; r(X_i) + \varepsilon_i \text{ for } i \in \mathbb{Z}$$

where the response Y is a real valued variable, and when the explanatory variable X belongs to some semi-metric space (H, d). The error process $(\varepsilon_i)_{i\in\mathbb{Z}}$ is assumed to be a centered second order stationary with unit variance, such that :

$$I\!E\,(\varepsilon_i\varepsilon_j) = f_\varepsilon(i-j) \text{ for } i,\, j \in \mathbb{Z}$$

Let (X, Y) be a random vector valued in $H \times \mathbb{R}$ with $\mathbb{E}(Y) < \infty$. So, for $x \in H$ the regression function is defined by :

$$r(x) = \mathbb{E}\left(Y|X = x\right)$$

The goal of this work is to estimate the unknown operator r. For this aim, we use the Nadaraya-Watson kernel estimator (see Ferraty and Vieu (2004))

$$\widehat{r}_h(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i\, K\left(\frac{d(x,X_i)}{h}\right)}{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)} & \text{if } K\left(\frac{d(x,X_i)}{h}\right) \neq 0 \\ 0 & \text{if } K\left(\frac{d(x,X_i)}{h}\right) = 0 \end{cases}$$

where K is a real valued function defined on $\mathbb{R}^+$ and $h = h(n)$ is the bandwidth, such that : $h \in \mathbb{R}^+$ and $\lim\limits_{n\to+\infty} h = 0$.

Results

We give the set of conditions that are necessary for the statement of our results.

– **About the kernel :**

$$\text{we assume that } K \text{ is strictly decreasing and there exist some positive numbers } a, b \text{ such that : } a\, \mathbb{1}_{[0,1]}(x) \leq K(x) \leq b\, \mathbb{1}_{[0,1]}, \text{ for } x \in H \qquad (5)$$

– **About the concentration of X :** we assume that the probability distribution of the functional variable X can be written as :

$$\text{for } x \in H,\ \P\left(X \in \mathcal{B}(x,h)\right) \;=\; C_x\varphi(h) + O\left(\varphi(h)\right), \qquad (6)$$

with $\sup\limits_{x\in S} C_x < \infty$ for some compact set $S \subset H$ and where $\mathcal{B}(x,h)$ denotes the closed ball of center x and radius h, and $\varphi(h)$ is a positive function such that the following conditions are satisfied :

$$\lim_{t\to 0} \varphi(t) = 0 \text{ and } \lim_{n\to\infty} n\,\varphi(h) = \infty \qquad (7)$$

– **About the regression :**

$$\exists C < \infty, \exists \beta > 0, \text{ such that}, \forall x, y \in H,\ |r(x) - r(y)| \leq \mathcal{C}\,(x,y)^{\beta} \qquad (8)$$

for some positive generic constant $\mathcal{C}$.

– **About the moments :**

$$\exists v \geq 2,\ \exists C_v > 0 \text{ such that } \mathbb{E}(Y^v|X) \leq C_v \qquad (9)$$

Théorème 1 *Under the conditions (5), (6), (7),(7 (8) and (9), we have almost surely :*

$$\begin{aligned}\sup_{x\in S}|\widehat{r}_h(x)-r(x)| &= O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\ln(n)}{n\varphi(h)}}\right)\\ &+O\left(\frac{\left(\ln(n)\sum_{i\neq j}|f_{\varepsilon}(i-j)|\right)^{1/2}}{n}\right)\end{aligned}\tag{10}$$

Remark 1 *Under conditions of Theorem 1 we have :*

$$\begin{aligned}\sup_{x\in S}|\widehat{r}_h(x)-r(x)| &= O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\ln(n)}{n\varphi(h)}}\right)\\ &+ \quad O\left(\sqrt{\frac{a_n\,\ln(n)}{n}}\right)+O(\frac{\ln(n)^{1/2}\,\alpha(a_n)^{1/2-1/v}}{\varphi(h)^{1-1/v}})\end{aligned}\tag{11}$$

for some sequence a_n such that : $a_n = o(n)$ and $\lim_{n\to+\infty} a_n = +\infty$.

Corollaire 1 *Under conditions of Theorem 1 and if $f_{\varepsilon}(j) = \mathcal{C}\,|j|^{-\gamma}$ and $0 < \gamma \leq 1$, then*

$$\begin{aligned}\sup_{x\in S}|\widehat{r}_h(x)-r(x)| &= O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\ln(n)}{n\varphi(h)}}\right)\\ &+ \quad O\left(\frac{\ln(n)^{1/2}}{n^{\gamma/2}}\right)\end{aligned}$$

Corollaire 2 *Under conditions of Theorem 1 if the data are independent identically distributed and come from some strongly mixing process, and that the mixing coefficients are such that :*

$$\exists q > 1,\ \alpha(n) \leq c\,n^{-q},\ \textit{for some positive constant } c. \tag{12}$$

then

$$\sup_{x\in S}|\widehat{r}_h(x)-r(x)| = O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\ln(n)}{n\varphi(h)}}\right)+O\left(\sqrt{\frac{\xi(h)}{\varphi(h)^2}\left[\frac{n}{\xi(h)}\right]^{s}\frac{\ln(n)}{n}}\right)$$

Remarks

– The results of convergence in the mean square sense will also be exposed.
– Some simulation studies will be presented in order to study the performance of the nonparametric estimators with long range dependent errors.

References

J. Beran (1992). Statistical methods for data with long range dependence. *J. Statist. Sci*, **7**, 404-420.

O. Cappe and E. Moulines and C. Pesquet and A. Petropulu and X. Yang (2002). Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Processing Magazing*, **3**, 14-27.

S. Csörgo and J. Mielniczuk (1995). Nonparametric regression under long-range normal errors. *Annals of Statistics*, **23**, 1000-1014.

G. Estévez and P. Vieu (2003). Nonparametric estimation under long memory dependence. *J. Nonparametric Statistics*, **15**, 535-551.

F. Ferraty and P. Vieu (2002). *Statistique fonctionnelle, modèle non paramétrique de régression*. Notes de cours de DEA (Univ. Paul Sabatier).

F. Ferraty and P. Vieu (2004). Nonparametric models for functional data, with application in regression method. *J. Nonparametric Statistics*, **16**, 111-125.

P. Hall and J. D. Hart (1990). Nonparametric regression with long range dependence. *J. Stoch. Proc. and Their Appli.*, **36**, 339-351.

W. Härdle (1989). *Applied nonparametric regression*. Cambridge, University Press.

J. Hart (1991). Kernel regression with time series errors. *J. of the Royal Statistical Society – Series B*, **53**, 173-187.

E. Masry (2005). Nonparametric regression estimation for dependent functional data asymptotic normality. *J. Stoch. Proc. and Their Appli.*, **115**, 155-177.

E. Masry and J. Mielniczuk (1999). Local linear regression estimation for time series. *J. Stoch. Proc. and Their Appli.*, **82**, 173-195.

Analyse non-paramétrique pour données fonctionnelles

Ali LAKSACI*, Frédéric FERRATY et Philippe VIEU

* Adresses pour correspondance :
Laboratoire de Statistique et Probabilités
Université Paul. Sabatier
Toulouse, France
et
Université Djillali Liabes
Sidi Bel Abbes, Algérie

e-mail : laksaci@cict.fr ou laksaci@yahoo.fr

Résumé

On se propose d'étudier la distribution d'une variable aléatoire réelle conditionnée par une variable fonctionnelle. L'objectif est l'estimation de la fonction de répartition conditionnelle, de la densité conditionnelle et de ses dérivées par la méthode du noyau. On établit la convergence presque complète de ces estimateurs et on applique ces résultats pour estimer le mode conditionnel et les quantiles conditionnels.

Notre étude met en évidence le phénomène de concentration de la mesure de probabilité de la variable fonctionnelle sur des petites boules. Ainsi, en utilisant les nombreux résultats récents en théorie des probabilités sur les petites boules, on peut préciser nos résultats pour de nombreux processus à temps continus. Finalement notre approche a été mise en application sur quelques données réelles de type spectrométrique ou de pollution.

Cette présentation concernera les cadres de données fonctionnelles indépendantes ou fortement mélangeantes.

Les résultats et exemples présentés sont issus des deux articles ci-dessous :

Références

1. Ferraty, F. ; Laksaci, A. ; Vieu., Ph. Estimating some characteristics of the conditional distribution in nonparametric functional models. Stat. Inference Stoch. Process. (a apparaitre)
2. Ferraty, F. ; Laksaci, A. ; Vieu., Functional times series prediction via conditional mode. C. R., Math., Acad. Sci. Paris (a paraitre)

Arbres de régression fonctionnels. Construction et applications en océanologie

David NERINI*, Claude MANTE, Badih GHATTAS

* Adresse pour correspondance :
Centre d'Océanologie de Marseille, UMR LMGEM 6117 CNRS, Campus de Luminy, Case 901, 13288 MARSEILLE Cedex 09
e-mail : nerini@com.univ-mrs.fr

Résumé

En océanologie, l'étude des écosystèmes et des interactions avec leur environnement nécessite fréquemment un échantillonnage spatial et temporel à haute fréquence sur un grand nombre de variables. L'utilisation de méthodes statistiques de prévision de type *CART*, connues pour leur capacité à traiter des volumes de données importants et pour leur structure attractive en forme d'arbre, sont largement utilisées dans ce domaine [1][2]. Cependant, elles ne permettent pas de prendre en compte l'aspect fonctionnel d'un grand nombre de variables d'intérêt. Par exemple, un profil de température échantillonné sur la verticale d'une colonne d'eau consiste en une succession d'observations indicées selon leur position en profondeur. Dès lors, il est intéressant de considérer cette suite de points comme un échantillonnage ponctuel d'une courbe que l'on va chercher à estimer.

Nous proposons de généraliser la méthode de régression *CART* lorsque la variable à prévoir est fonctionnelle. Ce travail est illustré à partir de différentes études en océanographie, dans lesquelles l'objectif principal est d'expliquer la structure d'une variable fonctionnelle (courbe de salinité, courbe d'ozone atmosphérique sur une journée, spectre de taille d'une communauté zooplanctonique, ...) en fonction d'un ensemble de variables explicatives réelles (température moyenne, quantité d'UV reçue, vitesse moyenne du vent, ...).

On dispose d'un échantillon d'apprentissage $L = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$ de taille n où les y_i sont les observations de la variable aléatoire expliquée Y à valeur dans un espace fonctionnel. Les $\mathbf{x}_i = (x_{i1}, .., x_{ip})$ constituent l'ensemble des observations de la variable aléatoire explicative $\mathbf{X} = (X_1, ..., X_p)$ à valeur

dans $\mathbb{R}^p$. A partir de la connaissance de L, on cherche à construire une procédure de régression :

$$f_L(\mathbf{x}) = E(Y/\mathbf{X} = \mathbf{x})$$

de la forme :

$$f_L(\mathbf{x}) = \sum_{j=1}^{q} f_j I(\mathbf{x} \in r_j)$$

où les f_j sont des fonctions, les r_j, polytopes de $\mathbb{R}^p$ dont les côtés sont parallèles aux axes, constituent une partition de $\mathbb{R}^p$, q est le nombre de partitions, inconnu a priori et I, la fonction indicatrice.

La construction du modèle est réalisée de manière séquentielle. Partant de l'échantillon L, l'espace $\mathbb{R}^p$ est initialement partitionné en deux régions dans lesquelles une estimation de Y est donnée. La partition retenue est celle qui maximise un critère construit à partir des observations de Y lorsqu'on passe de l'échantillon L à deux de ses partitions. Ce partitionnement dichotomique est reconduit dans les deux régions ainsi obtenues qui seront à leur tour partitionnées. Cette procédure est réitérée jusqu'à ce qu'une règle d'arrêt soit rencontrée. A la fin, l'espace $\mathbb{R}^p$ des variables de prévisions est partitionné en q regions $r_1, r_2, ..., r_q$. A chacune de ces régions est associée une cascade de règle de décisions emboîtées (les bords de chacun des r_j) construites sur les variables explicatives et une courbe prédite de Y. Le modèle peut alors être représenté sous la forme d'un arbre dichotomique.

La construction d'un arbre de régression dans le cas fonctionnel repose donc sur les trois étapes suivantes :

- La détermination du critère permettant de sélectionner les partitions,
- Le choix d'une règle d'affectation d'une fonction prédite dans chacune des partitions,
- Un critère d'arrêt dans la construction des partitions qui permettra de fixer la valeur de q.

Le critère proposé lorsque la variable réponse Y est à valeur dans un espace fonctionnel est une généralisation de la déviance, critère quadratique à la base de la construction des arbres de régression dans le cas univarié [1]. Sa construction repose sur l'hypothèse que chacune des observations de Y peut être décomposée sous la forme :

$$y_i(t) = c_{i0}\phi_0(t) + c_{i1}\phi_1(t) + ... + c_{ik}\phi_k(t) + \varepsilon(t)$$

où les $\phi_j, j = 0, .., k$ constituent une base de fonctions choisies à l'avance et les $c_{ij}, j = 0, ..., k$ les coefficients estimés par la projection de y_i dans la base des ϕ_j. La quantité ε est une variation résiduelle qui est considérée comme du bruit.

En s'appuyant sur les travaux de [3, 4, 5], nous discuterons successivement :
- des propriétés requises pour choisir un critère de partitionnement dans le cas multivarié
- de l'influence de la métrique liée au choix des ϕ_j sur la structure de l'arbre
- des choix possibles de représentations du modèle fonctionnel sous la forme d'un arbre.

Lorqu'un arbre a été construit, il est intéressant de mesurer ses performances. Or, un problème bien connu de ce type de modèle est leur instabilité : des changements mineurs dans l'échantillon d'apprentissage peuvent mener à des arbres de structures différentes et donc à des erreurs de prévisions importantes [6]. Nous montrons qu'il est possible de construire une version aggrégée de ces modèles fonctionnels par échantillonnage bootstrap de l'échantillon de construction [7]. Même si la structure arborée du modèle initial est perdue, il est possible, par ce moyen, d'améliorer de manière importante les erreurs de prévisions.

Références

[1] Breiman L., Friedman J.H., Olshen R., Stone C.J. (1984) *Classification And Regression Trees* , Wadsworth, Belmont CA.
[2] Nerini, D.,Durbec, J.P.,Mante, C.,Garcia, F.,Ghattas, B. (2000) Forecasting physicochemical variables by a classification tree method. Application to the Berre lagoon (South France). *Acta Biotheoretica*, 48 : 181-196.
[3] Segal M.R. (1992) Tree Structured Methods for Longitudinal Data *JASA* Vol. 87, N° 418, pp.407-418.
[4] Y. Yu, D. Lambert (1999) Fitting Trees to Functional Data : With an Application to Time-of-day Pattens. *Journal of Computational and Graphical Statistics*, Vol. 8, pp. 749-762.
[5] Zhang H. (1998) Classification Trees for Multiple Binary Responses *JASA* Vol. 93, N° 441, p180-193.
[6] Breiman L., (1996a) Heuristic of instability and stabilization in model selection, *The Annals of Statistics,* Vol 24, N°6, pp.2350-2383.
[7] Breiman L. (1996b) Bagging Predictors, *Machine Learning*, 24, pp. 123-140

.

Estimation non paramétrique de la régression pour des champs aléatoires à valeurs dans un espace métrique

Sophie NIANG

* Adresses pour correspondance :
Laboratoire GREMARS, Maison de la Recherche, Université Lille3
domaine du pont de bois, BP 60149, 59653 Villeneuve d'Ascq cedex

e-mail : sophie.dabo@univ-lille3.fr

Résumé

Nous étudions l'estimation non paramétrique de la régression dans le cas où la variable expliquée est un champ aléatoire réel tandis que la variable explicative est un champ aléatoire à valeurs dans un espace métrique. L'estimateur proposé est celui des k-points les plus proches. Nous donnons des résultats de convergence de l'estimateur.

Références

1. Tran, L.T., 1990. Kernel density estimation on random fields. J.Multivariate Anal. 34, 37-53.
2. Tran, L.T., Yakowitz, S., 1993. Nearest neighbor estimators for random fields. J.Multivariate Anal. 44, 23-46.

.

The ARHD model

Besnik PUMO*, André MAS

* Adresse pour correspondance :
Unité de Statistiques, Institut National d'Horticulture
2 rue Le Nôtre, 49045 Angers, France

e-mail : Besnik.Pumo@inh.fr

Résumé

Le modèle ARHD, introduit par Marion et Pumo (2004) est le processus $(X_i)_{i\in\mathbb{Z}}$ à valeurs dans l'espace de Sobolev $W = W^{2,1}[0,1]$ (voir par exemple Adams et Fournier) :

$$X_{i+1} = \phi(X_i) + \Psi\left(X_i^{'}\right) + \varepsilon_{i+1} \tag{13}$$

où ϕ et Ψ sont des opérateurs linéaires et $(\varepsilon_{i+1})_{i\in\mathbb{Z}}$ est un W bruit-blanc fort (voir Bosq 2000 par exemple). On suppose que ϕ est un opérateur compact de W dans W et Ψ un opérateur compact de $L = L^2[0,1]$ dans W. Soit D l'opérateur dérivé $Du = u^{'}$, $\langle u,v\rangle_W = \int_0^1 u(t)\,v(t)\,dt + \int_0^1 u^{'}(t)\,v^{'}(t)\,dt$ et $\langle u,v\rangle_L = \int_0^1 u(t)\,v(t)\,dt$. En nottant où $A = \phi + \Psi D$, on obtient la présentation$ARW(1)$

$$X_{i+1} = A(X_i) + \varepsilon_{i+1} \tag{14}$$

Cette équation a une solution unique stationnaire à valeurs dans W sous l'hypothèse :

$$\mathbf{H1} : \|A\| < 1 \tag{15}$$

$\|\cdot\|$ étant la norme usuelle des opérateurs. L'opérateur D étant borné dans notre contexte, on en déduit que le processus $\left(X_i^{'}\right)_{i\in\mathbb{Z}}$ à valeurs dans L est aussi strictement stationnaire .

Dans cet exposé nous présenterons la *méthode des moments* d'estimation des paramètres ϕ and Ψ et donnerons des résultats de convergence asymptotique sous la condition (précisée ci-dessous) que les deux paramètres soient identifiables. Pour prouver la convergence des estimateurs nous aurons de ce postulat technique :

$$\mathbf{H2} : \|X\|_W < +\infty \quad a.s. \tag{16}$$

qui a été par exemple proposé par Cardot, Ferraty, Sarda (1999) pour les mêmes raisons, et qui permet de simplifier les calculs.

Identifiablité du ARHD. Soit $\mathcal{C}_W$ (resp. $\mathcal{C}_{LW}$) l'espace des opérateurs compacts dans W (resp. de L dans W), $\mathcal{E} = \mathcal{C}_W \times \mathcal{C}_{LW}$,

$$\Gamma = E\left(X_0 \otimes_W X_0\right),\ \Gamma^{'} = E\left(X_0 \otimes_W X_0^{'}\right),$$

$$\Gamma^{'*} = E\left(X_0^{''} \otimes_L X_0\right),\ \Gamma^{``} = E\left(X_0^{'} \otimes_L X_0^{'}\right),$$

$$\Delta = E\left(X_0 \otimes_W X_1\right),\ \Delta^{'*} = E\left(X_0^{'} \otimes_L X_1\right).$$

Pour que les opérateurs soient uniquement définis il faut que :

$$\mathbf{H3} : \Gamma, \Gamma^{``} \text{ soient bijectives} \tag{17}$$

Estimation par la méthode des moments. Cette méthode est basée sur les équations $(\mathcal{S})$:

$$(\mathcal{S}) = \left\{ \begin{array}{l} \Delta = \phi\Gamma + \Psi\Gamma^{'} \\ \Delta^{'*} = \phi\Gamma^{'*} + \Psi\Gamma^{``} \end{array} \right. \tag{18}$$

On peut montrer que : *le couple* $(\phi, \Psi) \in \mathcal{E}$ *est identifiable par les équations des moments* (S) *ssi* $(\phi, \Psi) \notin \mathcal{N}$ *où :*

$$\mathcal{N} = \{(U, V) \in \mathcal{E} : U + VD = 0\}. \tag{19}$$

L'idée est alors d'estimer les opérateurs de covariance $\Gamma, \Gamma^{``}$ et covariance croisé $\Gamma^{'}, \Delta, \Delta^{'}$ par leurs estimateurs empiriques $\Gamma_n, \Gamma_n^{``}$ et covariance croisé $\Gamma_n^{'}, \Delta_n, \Delta_n^{'}$ pour en estimer ϕ et Ψ. Malheureusement les inverse $\Gamma, \Gamma^{"}$ ne sont pas bornés. L'idée est alors d'introduire des perturbations α et β tels que $\Gamma + \alpha I_W$, $\Gamma^{"} + \alpha I_L$,

$$S_\phi(\beta) = \Gamma - \Gamma^{'*}\left(\Gamma^{''} + \alpha I_L\right)^{-1}\Gamma^{'} + \beta I_W \tag{20}$$

$$S_\Psi(\beta) = \Gamma^{"} - \Gamma^{'}\left(\Gamma + \alpha I_W\right)^{-1}\Gamma^{'*} + \beta I_L \tag{21}$$

soient inversibles. Les équations :

$$(\mathcal{S}^{'}) = \left\{ \begin{array}{l} \Delta = \phi\left(\Gamma + \alpha I_W\right) + \Psi\Gamma^{'} \\ \Delta^{'*} = \phi\Gamma^{'*} + \Psi\left(\Gamma^{"} + \alpha I_L\right) \end{array} \right. \tag{22}$$

nous permet finalement d'obtenir les approximations :

$$\widetilde{\phi} = T_\phi[S_\phi(\beta)]^{-1} \tag{23}$$

$$\widetilde{\Psi} = T_\Psi[S_\Psi(\beta)]^{-1} \tag{24}$$

où :

$$\left\{ \begin{array}{l} T_\phi = \Delta - \Delta^{'}\left(\Gamma^{"} + \alpha I_L\right)^{-1}\Gamma^{'} \\ T_\Psi = \Delta^{'*} - \Delta\left(\Gamma + \alpha I_W\right)^{-1}\Gamma^{'*}. \end{array} \right. \tag{25}$$

De façon naturelle les expressions des estimateurs de ϕ et Ψ sont :

$$\left\{ \begin{array}{l} \phi_n = T_{n,\phi}[S_{n,\phi}(\beta_n)]^{-1} \\ \Psi_n = T_{n,\Psi}[S_{n,\phi}(\beta_n)]^{-1}. \end{array} \right. \tag{26}$$

où $T_{n,\phi}$, $S_{n,\phi}$, $T_{n,\phi}(\beta_n)$, $S_{n,\Psi}(\beta_n)$. sont les estimateurs empiriques des opérateurs respectifs. Le résultat principal de cette exposé est le suivant :

Sous les hypothèses $\mathbf{H1-3}$ *et si* $\alpha_n \to 0$, $\beta_n \to 0$ *avec* $\sqrt{n}\alpha_n^2\beta_n^2 \to +\infty$ *et* $\sqrt{\alpha_n}/\beta_n \to 0$,

$$\phi_n \to_{\mathbb{P}} \phi,$$

$$\Psi_n \to_{\mathbb{P}} \Psi.$$

Les conditions sont satisafites par exemple quand $\alpha_n = n^{-a}$ et $\beta_n = n^{-b}$ avec $b < a/2$ et $2b + 2a < 1/2$.

Quelques études numériques. Ces études concernent deux applications : Processus de Wong et ENSO (voir par exemple Besse et al, 2000). Les prédicteurs obtenus à l'aide de ARHD ont été comparé avec différentes méthodes de prédictions (Antoniadis et Sapatinas, 2001 ; Besse et al. 2000, Pumo, 1998).

Le processus de Wong est un processus Gaussien statonnaire d'ordre deux, à trajectoires 1 fois dérivables, défini pour $u \in R$ par :

$$\xi_u = \sqrt{3}\,\exp\left(-\sqrt{3}\,u\right)\int_0^{\exp\left(2u/\sqrt{3}\right)} W_s\,ds. \tag{27}$$

Un calcul direct montre qu'il accepte la présentation ARHD (à valeurs dans $W^{2,1}[0,\delta]$) avec $X_{i+1}(t) = \xi_{i\cdot\delta+t}$ pour $t \in]0,\delta]$.

Soit $(e_j, j \geq 0)$ la base de Fourier et $\mathbf{w} = \{e_0, [1 + 4j^2\pi^2/\delta^2]^{-1/2} \cdot e_{2j-1}, [1 + 4j^2\pi^2/\delta^2]^{-1/2} \cdot e_{2j}, j \geq 1\}$. Les estimateurs ARHD sont calculés en utilisant la propriété : $f \in W^{2,1}$, $f = \sum_{i=0,\infty} c_j e_j$ avec $c_j = \langle f, e_j\rangle_{L^2} \Rightarrow f^{'} = \sum_{i=0,\infty} c_j e_j^{'}$ et $f = \sum_{j=0,\infty} \langle f, w_j\rangle_W\, w_j$ où $\langle f, w_{2j-1}\rangle_W = [1 + 4j^2\pi^2/\delta^2]^{-1/2} \cdot \langle f, e_{2j-1}\rangle_L$.

Références

Adams R.A. and Fournier J.J.F., 2003. *Sobolev spaces*, Academic Press, 2nd ed.

Antoniadis A., Sapatinas T., 2003. Wavelet methods for continuous-time prediction using representations of autoregressive processes in Hilbert spaces, *J. Mult. Anal.,* 87, 133–158.

Besse, P., Cardot, H. and Stephenson, D., 2000. Autoregressive forecasting of some climatic variations, *Scand. J. Statist*, 27, 673-687.

Bosq, D., 1991. Modelization, nonparametric estimation and prediction for continuous time processes. *In : Roussas (Ed), Nato Asi Series C*, 335, 509-529.

Bosq, D., 2000. *Linear processes in function spaces*. Lectures notes in statistics. Springer Verlag.

Cardot H., Ferraty F., Sarda P., 1999. Functional linear model. *Statist. Probab. Lett.,* 45, 11-22.

Marion J.M., Pumo B., 2004. Comparaison des modèles ARH(1) et ARHD(1) sur des données physiologiques, *Annales de l'ISUP*, 48, 3, pp. 29-38.

Mas A., Menneteau L., 2003. Perturbation appraoch applied to the asymptotic study of random operators, *Progress in Probability*, 55, 127-134.

Mas A., Pumo B., 2005. The ARHD model, *Submitted*.

Pumo B., 1998. Prediction of continuous time process by C[0,1]-valued autoregressive process. *Stat. Infer. Stoch. Proc.*, 3, vol. 1, 297-309.

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

Plug-in bandwidth selection in nonparametric hazard estimation : Seismology applications

Alejandro QUINTELA del RIO

* Adresse pour correspondance :
Departamento de Matemáticas, Facultad de Informática, Campus de Elviña,
15071 A Coruña, Spain

e-mail : aqdr@lycos.es

Abstract

A seismic series is a set of earthquakes occurring in a given period of time in a given area. Earthquakes of a seismic series are considered as stochastic mathematical variables, belonging to a continuous space-time-energy medium with dimension 5 $(\Psi_i, \lambda_i, d_i, t_i.M_i)$, where Ψ_i and λ_i are the latitude and longitude of the epicenter, d_i the depth of the focus, t_i the origin time and M_i the magnitude (Udias and Rice, 1975). If the activity develops without abrupt changes, it is possible to know the structure that exists between the earthquakes (Torcal et al, 1999).

We center our attention in the one-dimensional series of times $\{t_i\}$, obviously when we have selected a particular geographical region and the other four parameters are perfectly detailed. If we consider the time series formed with the time intervals between consecutive earthquakes, that is $x_i = t_i - t_{i-1}$, the hazard function, or risk function, defined by

$$r(x) = \lim_{\Delta x \to 0} \frac{P\left(x \leq X < x + \Delta x \mid X \geq x\right)}{\Delta x} = \frac{f(x)}{1 - F(x)}. \tag{28}$$

measures the instantaneous risk at time x.

As in other nonparametric estimation settings (density, regression), is necessary to get a optimal bandwidth selection method to have good theoretical and practical properties of the kernel estimator. In Estévez and Quintela (1999), a cross-validation criterion is considered. This method consists in to choose the bandwidth h_{CV} that minimizes

$$CV_{l_n}(h) = \int r_h^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{\left(1 - F_h^{-i}(X_i)\right)\left(1 - F_n(X_i)\right)}, \tag{29}$$

where $F_n(\cdot)$ is the empirical distribution function of the data, and the function CV is constructed to mimic the global estimation error

$$MISE(h) = E\int \left(r_h(x) - r(x)\right)^2 dx. \tag{30}$$

The other functions used are

$$f_h^{-i}(x) = n_{l_n}^{-1} \sum_{|j-i|>l_n} \frac{1}{h} K\left(\frac{x - X_j}{h}\right) \tag{31}$$

and

$$F_h^{-i}(x) = n_{l_n}^{-1} \sum_{|j-i|>l_n} H\left(\frac{x - X_j}{h}\right), \tag{32}$$

that is, the kernel estimators of the density and distribution function, respectively, when we use in the estimation all the data except the closest points (in time) from X_i. Here l_n is a positive integer such that n_{l_n} satisfies $nn_{l_n} = \#\left\{(i,j)\,/\,|i-j| > l_n\right\}$. The election of l_n depends on the quantity of dependence between the data (e.g. seeing the autocorrelation of the data set).

In Estévez, Quintela and Vieu (2002), the bandwidth selected by this method is "penalized" by a quantity λ_n, that depends on the data, and it improves the results obtained by the cross-validation criterion. This is defined by

$$h_{CV}^{p} = h_{CV} + \lambda_n. \tag{33}$$

In practice, they check that a good result is to choose

$$\lambda_n = \left(0.8e^{7.9\widehat{\rho}} - 1\right) n^{-3/10} \frac{h_{CV}}{100}, \tag{34}$$

where $\widehat{\rho}$ is a consistent estimator of ρ (the autocorrelation of the data set), and h_{CV} is the cross-validation bandwidth with $l_n = 0$.

In this work, we study the plug-in method to select the smoothing parameter h. This method follows the asymptotic decomposition of the error (30) :

$$MISE(h) = C_1(nh)^{-1} + C_2(h^{2k}) + o(MISE(h)), \tag{35}$$

where C_1 and C_2 are constants that depends of the kernel and the hazard function. Minimizing the two first terms of this function, we obtain that the asymptotically optimal bandwidth has the form

$$h_{AMISE} = Cn^{-1/5}, \tag{36}$$

with C a constant with depends on the kernel K and the functions r, f and F. A plug-in bandwidth h_{PG} substitutes the unknown functions r, f and F by nonparametric estimates of them. Obviously, a nonparametric estimation of anyone of these functions has to make use of a pilot smoothing parameter as well.

In a simulation study, we check that smaller errors and much less sample variability can be reached, when we compare the plug-in bandwidth selection with the cross-validation one. Also, we can see that a good election for the pilot bandwidth can be done by means of the cross-validation one.

The theoretical results are applied to seismicity studies, through a real data set corresponding to a collection of seismic data of Spain, registered by the Instituto Geográfico Nacional (IGN) in the five last years (2000-2004, both included). In this catalogue, we have chosen two differentiated regions to compare their seismic risk, through the estimation of the hazard function : The Granada Basin (De Miguel et al., 1992) and the Triangle "Triacastela-Samos-Sarria" in the Galician Region (Estévez, Lorenzo and Quintela, 2002). The first area is situated in the central part of the Betic Cordilleras (southern Spain). It contains the Granada Basin and several mountains ranges around it. This is a geologically well studied region by different authors. The second region, at the northwest of Spain, is a geographic area in which the seismic activity increased considerably in the 1990's. Due to the mentioned increment of the seismic activity in this period, and, simultaneously, of the social alarm, the IGN increased the number of stations in this region, and therefore the number of available data is larger and these much more reliable.

References

De Miguel, F., Ibáñez, J., Alguacil, G., Canas, J., Vidal, F., Morales, J., Peña, J., Posadas, A. and Luzón, F., (1992), 1-18 Hz L_g attenuation in the Granada Basin (southern Spain). *Geophys. J. Int.*, **111**, 270-280.

Estévez, G. and Quintela, A., (1999), Nonparametric estimation of the hazard function under dependence conditions, *Communications in Statistics : Theory and Methods,* **28, 10**, 2297-2331.

Estévez, G., Lorenzo, H., and Quintela, A., (2002), Nonparametric analysis of the time structure of seismicity in a geographic region, *Annals of Geophysics,* **45**, 497-512.

Estévez, G., Quintela, A. and Vieu, P., (2002), Convergence rate for cross-validatory bandwidth in kernel hazard estimation from dependent samples, *Journal of Statistical Planning and Inference,* **104**, 1-30.

Quintela-del-Río, A., (2005), Plug-in bandwidth selection in kernel hazard estimation from dependent data. Preprint.

Torcal, F. , Posadas, A., Chica, M. and Serrano, I. (1999), Application of conditional geostatistical simulation to calculate the probability of occurrence of earthquakes belonging to a seismic series. *Geophysical Journal International*, **139**, 703-725.

Udias, A. and Rice, J. (1975), Statistical analysis of microearthquakes activity near San Andres Geophysical Observatory, Hollister, California, *Bulletin of the Seismological Society of America,* **65**, 809-828.

Quelques développements récents en Statistique Opératorielle

Yves ROMAIN

* Adresse pour correspondance :
Laboratoire de Statistique et Probabilités
UMR CNRS C5583, Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse cedex

e-mail : romain@cict.fr

Résumé

Dans cet exposé, nous nous proposons de mieux cerner la "Statistique opératorielle" à travers divers travaux récents et quelques ouvertures potentielles.

Ainsi, dans une première partie, nous insistons sur le caractère *"intrinsèque"* de la Statistique opératorielle : le langage/formalisme basé sur les opérateurs (où les projecteurs jouent un rôle important) permet d'appréhender les méthodes multidimensionnelles dans un cadre dégagé de contraintes de bases et de prendre en compte les divers problèmes liés à la dimension (qu'elle soit finie, "large" ou asymptotique, infinie). De plus, l'opération de *tensorisation* permet d'en élargir le champ des applications.

Dans une seconde partie, nous présentons trois exemples récents

- de travaux concernant *en amont* les propriétés d'opérateurs ([6]),

- de travaux sur l'*analyse canonique de deux sous-espaces relativement à un troisième* ([2]),[3]), et enfin,

- des études sur les p*roduits tensoriels de processus* stationnaires ([1]).

Enfin, en conclusion, on s'interesse à des domaines connexes dont le cadre est (naturellement ([5]) ou potentiellement [4]) opératoriel, ce qui permet d'envisager quelques investigations futures.

Références

1. Boudou, A. et Romain, Y., (2002), "*On spectral and random measures associated to continuous and discrete time processes"*, Stat. Proba. Letters 59 : 145-157.
2. Dauxois, J., Nkiet, G.M. et Romain, Y. (2004), *"Canonical analysis relative to a closed subspace",* Lin. Alg. Appl., Special Issue in Statistics, 388, 119-145.
3. Dauxois, J., Nkiet, G.M. et Romain, Y. (2004), *"Linear Relative Canonical Analysis. Asymptotic theory and some applications",* Ann. Inst. Stat. Maths., 56, n°2, 279-304.
4. Hyvarinen, A., Karhunen J. , et Oja, E., (2001), Independent Component Analysis., J. Wiley, NY.
5. Malley, J. D. et Hornstein, J., (1993), *"Quantum statistical inference*". Statist. Sci. 8, **4** 433-457.
6. Romain, Y., (2002), *"Perturbation of functional tensors with applications to covariance operators",* Stat. Proba. Letters 58 : 253-264.

Nonlinear Principal Component Analysis

Ernesto SALINELLI

* Adresse pour correspondance :
Dipartimento di Scienze Economiche e Metodi Quantitativi
Università del Piemonte Orientale "A. Avogadro"
Via Perrone, 18 - 28100 NOVARA, ITALIA

e-mail : ernesto.salinelli@eco.unipmn.it

Abstract

Linear Principal Component Analysis (LPCA) of a (real) random vector $\mathbf{X}$ of dimension p is a well-known statistical technique (see Pearson (1906) and Hotelling (1933)) used as a reduction tool in the sense that one looks for a linear transformed random vector $\mathbf{Y}$ of $\mathbf{X}$ of dimension $q < p$ explaining the most of the variability of $\mathbf{X}$.

Some authors in the past have pointed out the importance of considering transformations that depend on the covariability of the components X_j between their moments greater than the second one. This literature presents extensions of LPCA in several different directions, as the *polynomial PCs* of Gnanadesikan (1977), the *nonlinear principal components* of De Leeuw (1981) and Gifi (1990), the *additive principal component* of Donnell, Buja and Stuezle (1994).

In Salinelli (1998) and (2001) we introduce a notion of *nonlinear principal components* (NLPCs) extending the classical problem of maximizing the variance of a r.v. $\mathbf{X}$ to a function space more general than the one of linear functions, obtaining, from a mathematical point of view, a variational problem on an infinite dimensional Hilbert space.

In this talk we present the definition and the main results obtained on the existence and the properties of NLPCs, discussing their statistical significance and their relations with some of the existing alternative definitions.

In particular, we expose some results for absolutely continuous with positive

bounded densities and Gaussian r.v.'s. We show how the NLPCs can be considered a theoretically well-founded generalization of LPCs for several reasons : they solve a maximum variance problem, are invariant under orthogonal transformations, their variances are the eigenvalues of an opportune integral operator and the corresponding eigenfunctions form an orthonormal basis for the function space where the maximum variance problem is solved.

Finally, we present some more recent result on a study performed with Aldo Goia on the *marginal nonlinear principal components.*

References

De Leeuw, J.- Van Rijekevorsel, H.- Van der Wouden, H. (1981) Nonlinear Principal Components Analysis with B-Splines, *Methods of Operations Research*, 33, 379-393

Donnel, D.J.-Buja, A.-Stuetzle, W. (1994) Analysis of additive dependencies and concurtivities using smallest additive principal components, *Ann. Statist.* **22**, 1635-1673

Gifi, A. (1990) *Nonlinear Multivariate Analysis*, Wiley, New York

Gnanadesikan, R. (1977) *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, 24, 417-441, 498-520

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space, *Phil. Mag.* (6), 2, 559-572

Salinelli E., Nonlinear principal components I. Absolutely continuous random variables with positive bounded densities, *The Annals of Statistics*, Vol. 26, No. 2, 1998, 596-616

Salinelli E., Nonlinear Principal Components II : The Normal Distribution, WP n.7 del Dipartimento di Scienze Economiche e Metodi Quantitativi, Università del Piemonte Orientale "A. Avogadro", 2001