Groupe de travail STAPH: Statistique Fonctionnelle et Opératorielle 9

5èmes Journées STAPH de Lille, 21-22 Juin 2007

Coordinateurs
S. Dabo-Niang, E. Ould-Said, A. Boudou,
F. Ferraty, Y. Romain, P. Sarda, P. Vieu et S. Viguier-Pla

5èmes JOURNEES DE STATISTIQUE FONCTIONNELLE ET OPERATORIELLE

Université Charles De Gaulle (Lille 3) et Université du Littoral Côte d'Opale

Jeudi 21 et Vendredi 22 Juin 2007

Maison de la Recherche, Lille 3

La Statistique Fonctionnelle et Opératorielle occupe désormais une place importante dans la recherche en Statistique. Cet intérêt provient autant du large potentiel d'applications (imagerie, télédétection, météorologie, médecine, économie, ...) 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 en France. Ainsi, des chercheurs membres du LSP (Laboratoire de statistique et probabilités de Toulouse) animent sur ce thème le groupe de travail STAPH[1] dont le prolongement a été l'organisation de rencontres en juin 2002, juin 2003 et juin 2005 à l'Université Paul Sabatier et en juin 2006 à l'Université Pierre Mendès-France.

Après Toulouse et Grenoble, la cinquième édition des journées a lieu sous l'égide des Universités Charles de Gaulle (Lille 3) et de l'Université du Littoral Côte d'Opale et dans les locaux de la maison de la recherche de Lille 3. Elles seront une fois de plus l'occasion de mettre en lumière les différents domaines théoriques et appliqués de la Statistique Fonctionnelle et Opératorielle à travers des échanges entre chercheurs issus de laboratoires français et étrangers.

Enfin, nous tenons aussi d'ores et déjà à prendre rendez-vous pour la sixième édition de ces journées qui seront combinées avec le *1st International Workshop on Functional and Operatorial Statistics* qui aura lieu à Toulouse du 19 au 21 Juin 2008.[2]

[1] Toutes les informations concernant ces journées ainsi que sur l'ensemble des activités du groupe STAPH sont disponibles en ligne à l'adresse *http://www.lsp.ups-tlse.fr/staph/*

[2] Toutes les informations concernant ce workshop sont disponibles en ligne à l'adresse *http://www.lsp.ups-tlse.fr/staph/IWFOS2008/*

Comité Organisateur

- Ouafae Benrabah (Univ. Littoral Côte d'Opale)
- Nadia Bensaid (Univ. Lille 3, Lille)
- Laurence Broze (Univ. Lille 3, Lille)
- Sophie Dabo-Niang (Univ. Lille 3, Lille)
- Christian Francq (Univ. Lille 3, Lille)
- Elias Ould-Said (Univ. Littoral Côte d'Opale)
- Anne-Françoise Yao (Univ. Aix-Marseille 2)
- Jean-Michel Zakoain (Univ. Lille 3, Lille)

Comité Scientifique

Liste des conférenciers

Mohammed ATTOUCH (Univ. Sidi-Bel-Abbès, Algérie). *Asymptotic distribution of robust estimator for functional nonparametric models.* **attou-kadi@yahoo.fr**.

Hamdi RAISSI (Univ. Lille 3). *Testing the cointegration rank with the likelihood ratio test under uncorrelated but nonindependent errors assumption.* **hamdi.raissi@etu.univ-lille3.fr**.

Juan CUESTA-ALBERTOS (Univ. Cantabria, Espagne). *Random projections and goodness of fit tests for functional data.* **cuestaj@unican.es**.

Antonio CUEVAS (Univ. Autonome, Madrid). *Functional data analysis based on depth measures defined via projections.* **antonio.cuevas@uam.es**.

Aliou DIOP (Univ. Gastron Berger, Sénégal). *Estimateur Généralisé de Hill.* **alioudiop52@yahoo.fr**.

Manuel FEBRERO (Univ. Santiago de Compostela). *Outlier detection for functional data.* **mfebrero@usc.es**.

Aldo GOIA (Univ. Novara). *Some results on marginal nonlinear principal components.* **aldo.goia@eco.unipmn.it**.

Claude MANTE (Univ. Aix-Marseille 2). *Analyse en Composantes Principales de mesures : applications en Océanologie.* **Claude.Mante@com.univ-mrs.fr**.

André MAS (Univ., Montpellier 2). *ACP fonctionnelle locale et petites boules.* **mas@math.univ-montp2.fr**.

Paulo Eduardo OLIVEIRA (Univ. Coimbra, Portugal). *Asymptotics for kernel estimation with functional data.* **paulo@math.uc.pt**.

Juan Carlos PARDO FERNANDEZ (Univ. Santiago, Espagne). *Testing for the equality of regression curves with functional data.* **juancp@uvigo.es**.

Lubos PRCHAL (Charles Univ. Prague and Univ. Toulouse 3). *On testing equivalence of two ROC curves.* **prchal@cict.fr**.

Cristian PREDA (Univ. Lille 2). *PLS approach for discriminant analysis on functional data. Anticipated prediction.* **cristian.preda@univ-lille2.fr**.

Noureddine RHOMARI (Univ. Mohammed 2, Oujda, Maroc). *Inégalités maximales pour les sommes de vecteurs aléatoires banachiques dépendants et Applications.* **rhomari@fso.ump.ma**.

Fabrice ROSSI (INRIA, Paris). *Discrimination de fonctions par machines à vecteurs de support.* **Fabrice.Rossi@apiacoa.org**.

Ingrid VANKEILEGOM (Univ. Louvain, Belgique). *On the use of the bootstrap in nonparametric functional regression.* **vankeile@stat.ucl.ac.be**.

Céline VIAL (ENSAI, Rennes). *Assessing the finite dimensionality of functional data.* **celine.vial@univ-rennes1.fr**.

Qiwei YAO (London school of Economics, United Kingdom). *Spatial Smoothing in Relation to Nugget Effect.* **q.yao@lse.ac.uk**.

Résumés

- **Mohammed ATTOUCH (Univ. Sidi-Bel-Abbès, Algérie). Asymptotic distribution of robust estimator for functional nonparametric models**. Joint work with **Ali LAKSACI.**

  We propose a family of robust nonparametric estimators for regression function based on kernel methods. We establish the asymptotic normality of these estimators under the concentration properties on small balls of the probability measure of the functional regressors. A useful application to the prediction problem, to the discrimination in a semi-metric space and to the determination of confidence bands is given. In addition, to highlight the generality of our purpose and to emphasize on the role of each of our hypothesis, several special cases of our general condition are also discussed. Some simulations results are given to illustrate on finite samples the performance of our asymptotic normality. Finally, our method has been implemented and applied to some chemiometrical data.

- **Hamdi RAISSI (Univ. Lille 3). Testing the cointegration rank with the likelihood ratio test under uncorrelated but nonindependent errors assumption**.

  We study the asymptotic behaviour of the reduced rank estimator of the cointegrating space and adjustment space for vector error correction time series models with nonindependent innovations. It is shown that the distribution of the adjustment space can be quite different for models with iid innovations and models with nonindependent innovations. We show that the likelihood ratio test remains valid when the assumption of iid Gaussian errors is relaxed. Monte Carlo experiments illustrate the finite sample performance of the likelihood ratio test using various kinds of weak error processes.

- **Juan CUESTA-ALBERTOS (Univ. Cantabria, Espagne). Random projections and goodness of fit tests for functional data.**

  Counterexamples showing two different multidimensional distributions with a common marginal are very well known. The usual way to construct those counterexamples starts by fixing the marginal and, then, constructing two different distributions sharing this marginal.

  In [1] a different point of view is taken. There, the authors begin by having two multidimensional distributions P and Q, and, then, they consider the following problem: Given a continuous probability measure μ (for instance, gaussian), which is the μ-measure of the vectors, h, which satisfy that the (one-dimensional) marginals of P and Q along the line determined by h coincide? The answer is 1 if $P = Q$ and 0 if P and Q are different. Two sample goodness-of-fit tests follow straightforward from this result.

  Moreover, it was shown in [2] that this result can be extended to cover some families of distributions, thus providing ways to construct goodness-of-fit tests to those families. In particular, there it is shown that a distribution is gaussian if and only if almost every (one-dimensional) projection is gaussian.

  Results in [1] and [2] include the functional setting.

  In this talk I will comment those results and will present some applications to real data sets.

  References

  [1] Cuesta-Albertos, J.A., R. Fraiman and T. Ransford. (2007). A sharp form of the Cramér-Wold theorem. To appear in *J. Theoret. Probab.*

  [2] Cuesta-Albertos, J.A., E. del Barrio, R. Fraiman and C. Matrán. (2007). The random projection method in goodness of fit for functional data. To appear in *Comput. Statist. and Data Anal.*

- **Antonio CUEVAS (Univ. Autonome, Madrid). Functional data analysis based on depth measures defined via projections**. Joint work with **Ricardo FRAIMAN**.

  The duality arguments relying on the use of projections play an outstanding role in the probability theory for general (Banach) spaces. Likewise, this projection methodology can be of interest in some inference problems arising in functional data analysis.

  In particular, some notions of data depth, useful for classification and inference purposes, are defined in terms of one dimensional projections. Their properties of strong consistency and asymptotic normality are established. Some specific applications are discussed.

- **Aliou DIOP (Univ. Gastron Berger, Sénégal). Generalized Hill Estimator**. Joint work with **Gane Samb LO**.

  We introduce a statistical process depending on a continuous time parameter whose any margin can arise as a kernel estimator. We define for $\tau > 0$

  $$H^{\tau}_{k,\,n} = \sum_{i=1}^{k} (\frac{i}{k})^{\tau} \log \frac{X_{n-i+1,\,n}}{X_{n-i,\,n}}.$$

  This class of estimators is in its formulation a special case of the Kernel-type estimators proposed in Csörgő *et al.* (1985) with $K(u) = u^{\tau-1}\mathbf{1}_{\{0<u<1\}}$. Under conditions on the kernel $K(.)$, Csörgő *et al.* (1985) established weak consistency (under H1, H2, H3, H4) and asymptotic normality (under H5, H6, H7). The estimators that we propose have a kernel which does not satisfy the conditions H4, H6, H7 of Csörgő *et al.* (1985) when $\tau \leq 1/2$. In this paper, we establish the strong consistency of the proposed estimator when $\tau > 0$. Its asymptotic variance is given with respect to the value of the parameter τ. In some situations, the Generalized Hill's estimator (with $\tau = 1/2$) performs better than the Hill's estimator.

  References

  Csörgő, S. , Deheuvels, P. and Mason, D. M. (1985). Kernel estimates for the tail index of a distribution. *Ann. Statist.*, **13**, 1050-1077.

- **Manuel FEBRERO (Univ. Santiago de Compostela). Outlier detection for functional data.** Joint work with **Pedro GALEANO** and **Wenceslao GONZÁLEZ MANTEIGA**.

  Functional data arise nowadays in a great amount of scientific fields associated with monitoring process whose final outputs are samples of functions. As an example, several government agencies provides information in real-time about the level of a certain pollutant that can be considered like a trajectory along a specific period (say, one day). A considerable effort is being made in order to adapt the usual statistical methods for this kind of data (Ramsay & Silverman (1997), Ramsay & Silverman (2002), Ferraty & Vieu (2006)).

  In this work, different depth measures for functional data are considered. The concept of depth is related with the aim to find the center of a data cloud and it is the analogous of the mode (median) for multivariate data. A depth measure provides an ordering of every datum from "center" to outward, so this ordering can be used for outlier identification. A rigorous definition of outlier in functional settings has not be given. We consider that a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed. Therefore, we assume that the whole set of curves have been drawn from the same stochastic process, and curves not compatible with this assumption are outliers.

  In order to identify outliers in functional datasets we make use of functional depths and proceed as follows. If an outlier is in the dataset, the corresponding curve will have a significatively low depth. Therefore, a way to detect functional outliers is to look for the curves with smallest depth. Its number is frequently unknown. Therefore, we propose to detect outliers in a given functional sample using a nonparametric procedure based on functional depths. Some real examples will be provided for illustrations purposes.

  References

  Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* Springer-Verlag, New York.

  Ramsay, J. and Silverman, (1997). *Functional Data Analysis.* Springer-Verlag, New York.

  Ramsay, J. and Silverman, (2002). *Applied Functional Data Analysis.* Springer-Verlag, New York.

- **Aldo GOIA (Univ. Novara). Some results on marginal nonlinear principal components**. **aldo.goia@eco.unipmn.it**. Travail effectué en collaboration avec **Ernesto SALINELLI**.

  The problem of nonlinear extensions of the classical *linear principal component analysis* has been treated in a large literature in order to face nonlinear dependencies between variables. Salinelli (1998 and 2001) proposed a more general definition of nonlinear principal components and sketched the definition of *marginal nonlinear principal components* (MNLPC) from a theoretical point of view for uniform and normal distributions. The aim of the work is to analyze the definition of MNLPC, to illustrate its statistical and probabilistic foundation and its worthiness. We also introduce an estimator based on B-spline approximation. We focus on MNLPC since the analysis represents a first step in the direction of the more general nonlinear principal components defined in Salinelli (1998).
  Let $\mathbf{X} = [X_1, X_2, \ldots, X_q]^T$ be a continuous q-dimensional random vector (r.v.) with $\mathbb{E}[\mathbf{X}] = \mathbf{0}$, finite second moments $\mathbf{\Sigma_X} = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]$ and density $f_{\mathbf{X}}$ with support a domain (an open and connected set) $D \subseteq \mathbb{R}^q$. Denoting with $W_j^{1,2}$ the *weighted Sobolev space* of square integrable functions with square integrable first derivative, we introduce

  Definition 1. *The $j-$th marginal nonlinear principal component is defined as the $j-$th linear principal component (LPC) of the r.v.*

$$\Psi(\mathbf{X}) = [\psi_1(X_1), \psi_2(X_2), \ldots, \psi_n(X_q)]^T$$

  where each $\psi_j \in W_j^{1,2}$ is the solution of the maximization problem

$$\begin{cases} \max\limits_{\varphi} & \mathbb{E}\left[(\varphi(X_j))^2\right] \\ \text{sub} & \mathbb{E}\left[(\varphi'(X_j))^2\right] = 1 \end{cases}$$

  If it exists, the solution of each maximization problem in definition 1 is the dominant eigenfunction of the operator $\mathcal{D}_j$ defined by $\mathcal{D}_j(\varphi) = -\varphi'' - \varphi' f'_{X_j}/f_{X_j}$ associated to its first characteristic value λ_j. In our work we analyze some results on the existence and some properties of these eigenfunctions ψ_j. More precisely we study the existence problem in terms of the moment generating function of X_j, the symmetry of the ψ_j's, the monotonicy of the ψ_j's; we show as each ψ_j characterizes the marginal distribution of X_j;we relate our results with the literature concerning the so-called Chernoff inequality and finally we give some explicit examples of analitic computation of ψ_j's.

  The estimation problem involves a two step procedure: first we obtain an estimation $\widehat{\psi}_j$ of the transformations ψ_j and therefore we apply the classical LPC

algorithm to the vector $\widehat{\Psi}(\mathbf{X})$.
For the first step, suppose we have a sample $\{X_{j,i}\}_{i=1,2,\ldots,n}$ of i.i.d. random variables drawn from X_j, a random variable valued on the compact set I. Denoting with $\mathcal{S}_{k,d}$ the set of spline functions S_k defined on I of order d and having $k-1$ interior knots, we define our spline estimate of ψ_j as the function $\widehat{\psi}_j \in \mathcal{S}_{k,d}$ which solves the following maximization problem

$$\begin{cases} \max\limits_{v} \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(v\left(X_i\right)\right)^2 \\ \text{sub } \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(v'\left(X_i\right)\right)^2 = 1 \end{cases} \qquad v \in \mathcal{S}_{k,d}$$

and satisfies

$$\frac{1}{n}\sum_{i=1}^{n} v\left(X_i\right) = 0$$

Problems defined by both last equations can be converted into a finite dimensional eigenvalue problem known in numerical literature as Rayleig-Ritz method and that can be solved by using some available computer packages.
Under suitable hypotheses, we obtain some asymptotic results. Performances of the estimator are also shown by a simulation study.

References

Chernoff H. (1981). A note on an inequality involving the normal distribution, *The Annals of Probability*, 9 (3), 533-535.

Cuadras C. M. and Fortiana J. (1995). A countinuous metric scaling solution for a random variable, *Journal of Multivariate Analysis*, 52, 1-14.

De Boor C. (2001). *A practical guide to splines*. Sringer Verlag.

El Faouzi N.E. and Sarda P. (1999). Rates of convergence for spline estimates of additive principal components, *Journal of Multivariate Analysis*, 68, 120-137.

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

Salinelli E. (1998). Nonlinear principal components I. Absolutely continuous random variables with positive bounded densities, *The Annals of Statistics*, 26 (2), 596-616.

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

- **Claude MANTE (Univ. Aix-Marseille 2). Analyse en Composantes Principales de mesures : applications en Océanologie**.

  Considérons une famille $\{\nu_1, \cdots, \nu_N\}$ de mesures bornées signées (voir [1]) sur $[a, b]$, représentées par les poids

  $$\{F_i(t_k) := \nu_i([a, t_k]),\ i \in [1, N],\ k \in [1, p]\}$$

  associés aux points d'une grille d'échantillonnage T_p. Les courbes granulométriques rencontrées en Sciences de la Terre constituent un cas typique de telles données.

  Nous avons récemment proposé dans [2] une méthode exploratoire pour l'analyse de la famille de densités $\{d\nu_1/d\mu, \cdots, d\nu_N/d\mu\}$, où la "probabilité de référence" μ est définie sur $[a, b]$. Alors que la grille T_p dépend de l'appareil utilisé, μ dépend du point de vue adopté pour l'analyse. Nous avons montré que cette analyse se ramène à l'ACP usuelle des vecteurs $\{F_1, \cdots, F_N\}$ dans une métrique dépendant de μ et de T_p.

  L'exposé sera articulé autour de l'étude comparative de trois campagnes sédimentologiques menées dans l'Etang de Berre en 1974, 1992 et 1997. Alors que dans [2] μ était associée au transport sédimentaire (tension de fond), elle correspondra ici à un sédiment de référence particulier.

  Références

  [1] Halmos, P. R. (1950). *Measure theory*, van Nostrand, New York.

  [2] Manté, C., Yao, A.F. and Degiovanni, C. (2007). Principal component analysis of measures, with special emphasis on grain-size curves. *Computational Statistics & Data Analysis*, sous presse.

- **André MAS (Univ., Montpellier 2). ACP fonctionnelle locale et petites boules**.

  En statistique non paramétrique classique l'estimation d'une fonction inconnue en un point fixé fait généralement intervenir des méthodes dites "locales" : on ne garde que les observations autour du point en question ou on leur affecte une pondération qui decroît avec l'éloignement. Cette pondération passe le plus souvent par l'introduction de noyaux. Les polynômes locaux constituent un exemple classique d'utilisation des méthodes locales.

  Dans le cas des données fonctionnelles les auteurs s'intéressent depuis peu à l'implémentation de ces méthodes mais des résultats théoriques font encore défaut.

  On propose de définir l'ACP fonctionnelle locale commme une ACP effectuée sur des opérateurs de covariance locaux. L'étude asymptotique des opérateurs (théoriques et empiriques) est menée. Elle nécessite de disposer d'un cadre très précis de travail portant sur les probabilités de petites boules shiftées associées à la distribution commune des observations.

  Références

  Dauxois J., Pousse A., 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.

  Ferraty F., Mas A., Vieu, P. (2007). Advances in nonparametric regression for functional variables. To appear in *Australian and New-Zealand Journal of Statistics.*

  Li, W.V., Shao Q.M. (2001). *Gaussian processes : Inequalities, small ball probabilities and appli- cations.* Handbook of Statistics, **19**, 533-597.

  Meyer-Wolf E., Zeitouni O. (1993). The probability of small gaussian ellipsoïds. *Annals of Probability*, **21**(1), 14-24.

- **Paulo Eduardo OLIVEIRA (Univ. Coimbra, Portugal). Asymptotics for kernel estimation with functional data.**

We consider the asymptotics of kernel estimators for functional data trying to adapt the techniques used in finite dimensional setting. Thus, this is a purely theoretical look at the estimation properties, proving extensions of traditional techniques to a more general framework. The density representations are replaced by the so called *small ball probability* assumptions. These introduce a density functional and a volume parameter that mimic the behaviour in the traditional setting. With reasonable assumptions on the density functional and the volume parameter, we may prove a version of a Bochner Lemma, thus serving as a basis for the proof of asymptotic results for the estimation of the regression function. This approach, assuming some knowledge on the volume parameter, also describes the estimation of the density functional. The proofs, suggested by the approach to estimation in point processes context, follow along somewhat classical arguments, but still we can find results based on assumptions comparable, and in some cases somewhat better, to the ones based in different approaches.

References

Bensaïd, N. and Fabre, J.-P., 1998, Convergence de l'estimateur à noyau de dérivées de Radon-Nikodym générales dans le cas mélangeant, *Canad. J. Statist.* **26**, 267–282.

Ferraty, F. and Vieu, P., 2004, Nonparametric models for functional data, with application in regression, time series prediction and curve estimation, The International Conference on Recent Trends and Directions in Nonparametric Statistics, *J. Nonparametr. Stat.* **16**, 111–125.

Masry, E., 2005, Nonparametric regression estimation for dependent functional data: asymptotic normality, *Stochastic Process. Appl.* **115**, 155–177.

Jacob, P. and Oliveira, P. E., 1997, Kernel estimators of general Radon-Nikodym derivatives, *Statistics* **30**, 25–46.

- **Juan Carlos PARDO FERNANDEZ (Univ. Santiago, Espagne). Testing for the equality of regression curves with functional data**. Joint work with **Frédéric FERRATY** and **Philippe VIEU.**

  Regression models are used for describing the relationship between a response variable and a covariate. When two or more groups can be distinguished in the populations, it is of interest testing for the equality of the corresponding regression functions in order to check whether the effect of the covariate on the response variable is the same in all groups.

  In a general context, the statistical model can be described as follows. For $j = 1, \ldots, k$, let (X_j, Y_j) be independent random vectors, where Y_j represents a certain response variable related to the covariate X_j via the regression model $Y_j = m_j(X_j) + \varepsilon_j$, where m_j is the regression function and ε is the regression error. We are interested in testing the null hypothesis of equality of the regression functions $H_0 : m_1 = m_2 = \cdots = m_k$.

  In the first part of the talk, we will revise the literature about comparison of regression functions. In the first articles concerning this topic, many restrictions are imposed on the model, such as equal design points or homoscedasticity. More recently, several methods have been proposed to treat general situations (more that two populations, heteroscedasticity, censored data).

  In the second part of the talk, we will present a recent researching project which deals with the comparison of regression curves in a functional data setup: we consider regression models with functional covariates and scalar responses. A test statistic is proposed in order to check the equality of the regression functions. The proposed test statistic is proved to be a U-statistic, and its properties are investigated.

- **Lubos PRCHAL (Univ. Toulouse 3). On testing equivalence of two ROC curves.**

ROC curves, see [9] for the actual state of the art, are a popular and widely used diagnostic tool having arisen in the context of signal detection theory developed in the 1950s and 1960s by [4]. After a publication of the text by [12], they became popular in radiology and nowadays they enjoy broad applications in medicine due to [6] and [10]. They play an important role in evaluation of classification and prediction methods in the applied statistical domains such as data mining (see [5]), information retrieval (see [7]), or e.g. computational linguistics (see [1]).

Our aim is to propose a nonparametric statistical methodology for comparing ROC curves which enables an objective, "fully automatic" statistical inference. First, we recall a notion of ROC curves and discuss its empirical and kernel estimators. The main part is focused on testing the equivalence of two ROC curves. The proposed methodology is based on quantile processes ([3] and [11]) associated to the curves. Their asymptotic properties are studied and several ways to obtain the critical values, based on both the asymptotic theory and simulations, are discussed. The suggested test is compared by the means of simulation with some other existing methods (see eg. [2], [13] and [14]) and finally applied to linguistic data concerning the collocation extraction (see [8]).

Acknowledgement. The research was partially supported by the project No. 135007 of the Grant Agency of Charles University, by the grant No. 201/05/H007 of the Grant Agency of the Czech Republic and by the Research Project No. MSM 0021620839 of the Ministry of Education of the Czech Republic.

Références

[1] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Technical Report*, Madison Computer Science Department, University of Wisconsin.

[2] DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrika*, **44**, 837 – 846.

[3] Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.*, **2**, 267 – 277.

[4] Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, New York.

[5] Fawcett, T. (2003). ROC graphs: Notes and practical consideration for data mining researchers. *HP Technical Report*.

[6] Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crtitical Reviews in Diagnostic Imaging*, **29**, 307−335.

[7] Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 329−338, ACM Press, New York.

[8] Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. *In: Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions*, Sydney, Australia, July 2006.

[9] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

[10] Shapiro, D. E. (1999). The interpretation of diagnostic tests. *Stat. Methods Med. Res.*, **8**, 113−134.

[11] Shorack, R. G. and Wellner, J. A. (1986). *Empirical Processes with Application to Statistics*. Wiley, New York.

[12] Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

[13] Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing reciever operating characteristic curves from a paired experiment. *Biometrika*, **83**, 835−848.

[14] Zhou, X. H., McClish, D. K. and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

- **Cristian PREDA (Univ. Lille 2). PLS approach for discriminant analysis on functional data. Anticipated prediction**.

  Partial least squares (PLS) approach is proposed for linear discriminant analysis (LDA) when predictors are data of functional type (curves). Based on the equivalence between LDA and the multiple linear regression (binary response) and LDA and the canonical correlation analysis (more than two groups), the PLS regression on functional data is used to estimate the discriminant coefficient functions. Anticipated prediction aspect is considered. A simulation study as well as an application to kneading data compare the PLS model results with those given by other methods.

- **Noureddine RHOMARI (Univ. Mohammed 2, Oujda, Maroc). Inégalités maximales pour les sommes de vecteurs aléatoires banachiques dépendants et Applications**.

Nous établissons, dans ce travail, des inégalités maximales, pour les sommes partielles de vecteurs aléatoires dépendants prennent leurs valeurs dans des espaces de Hilbert ou de Banach séparables, de dimensions finies ou infinies. Nous considérons deux types de dépendances, le mélange fort α et l'absolue régularité β. Ces inégalités sont pratiquement les mêmes que dans le cas réel dépendant.

Nous terminons par quelques applications à la loi forte des grands nombre et à la loi du logarithme itéré bornée pour des processus absolument réguliers hilbertiens ou banachiques, sous une condition faible sur le mélange.

Comme cas particuliers nous appliquons ces derniers résultats à l'estimation de l'opérateur de covariance de la marge d'un processus aléatoire, banachique ou hilbertien, absolument régulier (β-mélangeant), ainsi qu'à l'estimation de ses valeurs propres. Nous montrons aussi la forte consistance dans L^2 de l'estimateur récursif à noyau de la densité de probabilité, avec vitesse optimale, sous de faibles conditions de mélange d'absolue régularité (une décroissance logarithmique suffit pour la convergence p.s., et la vitesse est optimale si le mélange est sommable).

On obtient par exemple pour des vecteurs aléatoires centrés Y_i à valeurs dans un espace de Hilbert: si pour $1 \leq p \leq n/2$, $1 \leq i \leq 2r$, $\|Y_{(i-1)p+1} + \ldots + Y_{ip}\| \leq pM$, p.s., et $E\|Y_{(i-1)p+1} + \ldots + Y_{ip}\|^2 \leq p\sigma^2$ on a alors pour tout $\varepsilon > 0$

$$P\left(\max_{1\leq k\leq n}\left\|\sum_{t=1}^{k} Y_t\right\| \geq n\varepsilon + \left[\frac{p}{2}\right]_e M\right) \leq 4\exp\left(-\frac{n\varepsilon^2}{4[(1+2p/n)\sigma^2 + pM\varepsilon/3]}\right) + \left(\frac{n}{p}+2\right)\beta(p).$$

où $\beta(\cdot)$ est le coefficient de mélange régulier.

Nous montrons par exemple que la LFGN ait lieu dans des Banach de type $1 < \tau \leq 2$ dès que $\beta(n) = O((\log n)^{-1}(\log\log n)^{-c})$, avec $c > 2$. Et pour les v.a. hilbertiens on obtient la LIL bornée dès que $\beta(n) = O(n^{-a}\log^{-b} n \log_2^{-c} n)$ avec ($a > 1$ et $b, c \in \mathbb{R}$) ou ($a = 1$, $b > 1$ et $c \in \mathbb{R}$) ou ($a = b = 1$ et $c > 2$) et on a

$$\limsup_n \frac{1}{\sqrt{n\log\log n}} \max_{1\leq k\leq n}\left\|\sum_{t=1}^{k} Y_t\right\| \leq 2M\sqrt{1+10\sum_i \beta(i)}, \quad \text{p.s.}$$

L'estimateur récursif à noyau de la densité de probabilité est fortement consistant dans L^2, avec vitesse optimale, sous de faibles conditions de mélange d'absolue régularité (une décroissance logarithmique suffit pour la convergence p.s., et la vitesse est optimale si le mélange est sommable): pour $\hat{f}_n(x) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{h_t^d} K((x - X_t)/h_t)$, pour $x \in \mathbb{R}^d$, on a la borne suivante dans L^2,

$$\limsup_n \sqrt{\frac{nh_n^d}{\log\log n}} \|\hat{f}_n - E\hat{f}_n\|_2 \leq 2\|K\|_2 \sqrt{1 + 10\sum_i \beta(i)}, \quad \text{p.s.}$$

Références

[1] Bosq, D. (2000). *Linear Processes in Function Spaces. Theory and applications.* Lecture notes in Statistics. Springer.

[2] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces.* Springer-Verlag, New York.

[3] Pinelis, I.F. (1990). Inequalities for distribution of sums of independent random vectors and their application to estimating density. *Theory Probab. Appl.*, **35**, 605-607.

[4] Pinelis, I.F. and Sakhanenko, A.I. (1985). Remarks on inequalities for large deviation probabilities. *Theory Probab. Appl.*, **30**, 143-148.

[5] Rhomari, N. (2002). Approximation et inégalités exponentielles pour les sommes de vecteurs aléatoires dépendants. *C. R. Acad. Sci. Paris, Ser. I*, **334**, 149-154.

[6] Rio, E. (1995). A maximal inequality and dependent marcinkiewicz-zygmund strong laws. *Ann. Probab.*, **23**, 918-937.

- **Fabrice ROSSI (INRIA, Paris). Discrimination de fonctions par machines à vecteurs de support.** Travail en collaboration avec **Nathalie VILLA**

  Une machine à vecteurs de support (MVS, [6]) est un outil de discrimination basé sur la maximisation de la marge d'un séparateur affine : les données à classer sont envoyées dans un espace de Hilbert à noyau reproduisant (RKHS) dans lequel on choisit un séparateur affine en minimisant un compromis entre les erreurs de classement du séparateur et la norme du vecteur normal qui le définit. Le passage de l'espace de départ au Hilbert est réalisé de manière implicite en s'appuyant sur son noyau, dont la donnée est suffisante pour construire la MVS.

  Dans le cas de données fonctionnelles, le passage par un espace à noyau reproduisant peut sembler superflu, deux classes de fonctions étant généralement linéairement séparable dans un espace de dimension infinie. Cependant, une MVS ainsi construite s'appuie sur une régularisation de type *ridge* dont l'efficacité dans un cadre fonctionnel est assez limitée ([4]).

  Il est donc naturel d'étudier les noyaux adaptés aux données fonctionnelles. On étudie deux types de noyaux. Le premier consiste en la combinaison d'un noyau classique de MVS (par exemple le noyau Gaussien) avec un opérateur de projection sur une base tronquée : en s'inspirant de [2], on suppose que les fonctions observées sont éléments d'un Hilbert dont on se donne une base $\{\Psi_j\}_{j\geq 1}$. En utilisant une méthode de validation, on construit ainsi une MVS classique sur les d premières coordonnées sur la base des fonctions observées. Tous les paramètres de la MVS (d inclus) sont choisis automatiquement à partir des données, de façon consistante (l'erreur de la MVS converge avec le nombre de fonctions observées vers l'erreur bayésienne optimale, voir [5]).

  Le deuxième type de noyau est adapté aux données fonctionnelles régulières et discrétisées. On considère un opérateur différentiel $L = D^m + \sum_{j=0}^{m-1} a_j D^j$ définit sur l'espace de Sobolev $\mathcal{H}^m([0,1])$, qui conduit à un RKHS $\mathcal{H}_1$ (sous-espace de $\mathcal{H}^m$) dans lequel le produit scalaire est obtenu à partir de L ($\langle u, v\rangle_1 = \langle Lu, Lv\rangle_{L^2}$) (voir [1]). On suppose alors les fonctions observées à valeurs dans $\mathcal{H}_1$ et discrétisées en $t_1, \ldots, t_d$. À chaque fonction x, on associe son interpolation L-spline dans $\mathcal{H}_1$, qui coïncide exactement avec x en $t_1, \ldots, t_d$ et qui est de norme minimale. L'intérêt d'une MVS construite dans $\mathcal{H}_1$ de cette manière est que le noyau utilisé s'appuie sur l'opérateur L et permet donc de considérer les dérivées des fonctions observées, plutôt que les fonctions elles-mêmes, ce qui est fructueux en pratique (voir [5]). Or, on peut construire directement le noyau à partir des valeurs des fonctions observées en $t_1, \ldots, t_d$, sans passer par le calcul des fonctions d'interpolation. On montre de plus qu'une MVS obtenue de cette façon est consistante (voir [7]).

Références

[1] Besse, P. and Ramsay, J.O. (1986). Principal component analysis of sampled curves. *Psychometrica*, **51**, 285-311.

[2] Biau, G., Bunea, F. and Wegkamp, M. (2005). Functional Classification in Hilbert Spaces. *IEEE Transactions on Information Theory*, **51**, 2163-2172.

[3] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*, Cambridge University Press, UK.

[4] Hastie, T. and Buja, A. and Tibshirani, R. (1995). Penalized Discriminant Analysis. *Annals of Statistics*, **23**, 73-102.

[5] Rossi, F. and Villa, N. (2006). Support Vector Machine For Functional Data Classification. *Neurocomputing*, **69**, 730-742.

[6] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*,Springer Verlag New York.

[7] Villa, N. and Rossi, F. (2006). Un résultat de consistance pour des SVM fonctionnels par interpolation spline. *Comptes Rendus Mathématiques*, **343**(8), 555-560.

- **Ingrid VANKEILEGOM (Univ. Louvain, Belgique). On the use of the bootstrap in nonparametric functional regression.**

  We consider the functional regression model $Y = r(X) + \varepsilon$, where the response Y is univariate, X is a functional covariate, and the error ε satisfies $E(\varepsilon) = 0$. For this model, the asymptotic normality of a nonparametric kernel estimator $\hat{r}(\cdot)$ of $r(\cdot)$ has been proved [see Ferraty et al. (2007)]. In order to use this result for the construction of confidence intervals for $r(x)$ or for the selection of the smoothing parameter of the estimator $\hat{r}(x)$, the asymptotic variance of $\hat{r}(x)$ needs to be estimated. To circumvent the estimation of this variance, which is a difficult task, we study two bootstrap procedures to approximate the distribution of $\hat{r}(\cdot)$. Both a naive and a wild bootstrap procedure are studied, and their asymptotic validity is proved. The obtained consistency results are also tested via a simulation study, and a real data set is analyzed using these bootstrap procedures.

  Références

  Ferraty F., Mas A., Vieu, P. (2007). Advances in nonparametric regression for functional variables. *Australian and New-Zealand Journal of Statistics.*

- **Céline VIAL (ENSAI, Rennes). Assessing the finite dimensionality of functional data**. Travail en collaboration avec **Peter HALL**

  If a problem in functional data analysis is low-dimensional then the methodology for its solution can often be reduced to relatively conventional techniques in multivariate analysis. Hence, there is intrinsic interest in assessing the finite-dimensionality of functional data. We show that this problem has several unique features. From some viewpoints the problem is trivial, in the sense that continuously-distributed functional data which are exactly finite-dimensional are immediately recognisable as such, if the sample size is sufficiently large. However, in practice, functional data are almost always observed with noise, for example resulting from rounding or experimental error. Then the problem is almost insolubly difficult. In such cases a part of the average noise variance is confounded with the true signal, and is not identifiable. However, it is possible to define the unconfounded part of the noise variance. This represents the best possible lower bound to all potential values of average noise variance, and is estimable in low-noise settings. Moreover, bootstrap methods can be used to describe the reliability of estimates of unconfounded noise variance, under the assumption that the signal is finite-dimensional. Motivated by these ideas, we suggest techniques for assessing the

finiteness of dimensionality. In particular, we show how to construct a critical point $\hat{v}_q$ such that, if the distribution of our functional data has fewer than $q-1$ degrees of freedom, then we should be prepared to assume that the average variance of the added noise is at least $\hat{v}_q$. If this level seems too high then we must conclude that the dimension is at least $q-1$. We show that simpler, more conventional techniques, based on hypothesis testing, are generally not effective.

- **Qiwei YAO (London school of Economics, United Kingdom). Spatial Smoothing in Relation to Nugget Effect**.

  For spatio-temporal regression models with observations taken regularly in time but irregularly over space, we investigate the effect of spatial smoothing on the reduction of variance in estimating both parametric and nonparametric regression functions. The processes concerned are stationary in time but may be nonstationary over space. Our study indicates that the existence of the so-called nugget effect in either regressor process or noise process guarantees that spatial smoothing reduces the estimation variance. In particular the nugget effect in regressor process may lead to a faster convergence rate in estimating nonparametric regression functions.