# 4èmes Journées de Statistique Fonctionnelle et Opératorielle

Grenoble, 15-16 Juin 2006

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

## TABLE DES MATIERES

**Présentation des journées**. 5

**Samir BENAISSA et al.** : Prédiction fonctionnelle dans les inclusions différentielles stochastiques : Théorie et application. 11

**Delphine BLANKE et al.** : Estimation du nombre de dérivées d'un processus gaussien. 15

**Hervé CARDOT** : Conditional functional principal component analysis. 17

**Christophe CRAMBES et al.** : Smoothing Splines Estimators in Functional Linear Regression with Errors-in-Variables. 23

**Antonio CUEVAS et al.** : On classification methods with infinite-dimensional data. 29

**Pedro DELICADO** : Functional ANOVA when data are a weighted sample of density functions. 35

**Laurent DELSOL** : Normalité asymptotique de la régression sur variable fonctionnelle avec application à la construction d'intervalles de confiance. 39

**M'hamed EZZAHRIOUI et al.** : Asymptotic normality of conditional quantile in the normed space under $\alpha$-mixing hypothesis. 45

**M. FEBRERO et al.** : Some ideas on depth for clustering with functional data. 47

**Arnaud GUYADER et al.** : Classification par les plus proches voisins en dimension infinie. 53

**Aloïs KNEIP** : Common functional principal components analysis. 55

**Abbes RABHI et al.** : Estimation of conditional distribution and conditional hazard function. 57

**Jérôme SARACCO et al.** : Une méthode SIR multivariée en présence de covariables qualitatives. 67

**Yingcai SU** : Gene Expression Analysis between *Vitis vinifera* and the Disease-Resistant American Grapevine *Vitis aestivalis*. 73

**Anne-Françoise YAO** : Estimation du modèle de réduction de dimension fonctionnelle. 75

4

.

# Quatrièmes Journées de Statistique Fonctionnelle et Opératorielle

## Université Pierre Mendès France
## Maison des Sciences de l'Homme (MSH-Alpes)
## Grenoble, 15-16 Juin 2006

La Statistique Fonctionnelle et Opératorielle connaît un essor à l'échelle nationale et internationale. Ces journées sont la quatrième édition destinée à promouvoir cette discipline scientifique grâce à la rencontre entre chercheurs issus de laboratoires français et étrangers. L'accent est principalement mis sur une diversité d'approches balayant un champ large de la théorie aux applications : variables et modèles fonctionnels, statistique des opérateurs, statistique des processus, ... Les jeunes statisticiens ont été particulièrement encouragés à contribuer à cette manifestation.

Après l'organisation des trois premières éditions de ces journées à Toulouse par le groupe STAPH, elles ont lieu cette année à Grenoble dans les locaux et sous l'égide du LabSad. Depuis leur création, ces journées associent traditionnellement dans leurs travaux et dans leur organisation divers organismes de recherche : le laboratoire de Statistique et Probabilités de l'Université Paul Sabatier et le département BIA-INRA de Toulouse, l'Université de Sidi-bel-Abbès et le groupe SF&S (Statistique Fonctionnelle & Sondages) de Grenoble. Pour cette quatrième édition, d'autres organismes sont associés : les Universités de Bordeaux, Montpellier, du Littoral, de Lille 3, du Missouri (USA), de Novara (Italie), de Bonn (Allemagne), l'Université USTHB d'Alger (Algérie), les Universités Autonoma de Madrid, de Santiago de Compostela et Politècnica de Catalunya en Espagne, ainsi que l'ENESAD-INRA de Dijon.

En remerciant tous les participants,

Les comités d'organisation et scientifique

Grenoble, Juin 2006

# Comité Scientifique

- Karim Benhenni (Univ. Pierre Mendès-France, Grenoble)
- Alain Boudou (Univ. Paul Sabatier, Toulouse)
- Kamel Boukhetala (USTHB Alger, Algérie)
- Antonio Cuevas (Univ. Autonoma de Madrid, Espagne)
- Sophie Dabo-Niang (Univ. Lille 3)
- Frédéric Ferraty (Univ. Paul Sabatier/Univ. Le Mirail, Toulouse)
- Elias Ould-Said (Univ. Littoral)
- Aldo Goia (Univ. Novara, Italie)
- Mustapha Rachdi (Univ. Pierre Mendès-France, Grenoble)
- Yves Romain (Univ. Paul Sabatier, Toulouse)
- Pascal Sarda (Univ. Paul Sabatier/Univ. Le Mirail, Toulouse)
- Philippe Vieu (Univ. Paul Sabatier, Toulouse)
- Sylvie Viguier-Pla (Univ. Perpignan)
- Abderrahmane Yousfate (Univ. Sidi-Bel-Abbès, Algérie)

# Comité Local d'Organisation

- Taghi Barumandzadeh (Univ. Pierre Mendès-France, Grenoble),
- Karim Benhenni (Univ. Pierre Mendès-France, Grenoble),
- Mohamed El Methni (Univ. Pierre Mendès-France, Grenoble),
- Philippe Garat (Univ. Pierre Mendès-France, IUT 2, Grenoble),
- Iyadh Gacem (Univ. Pierre Mendès-France, Grenoble),
- Sonia Heldi-Griche (Univ. Pierre Mendès-France, Grenoble),
- Michel Lejeune (Univ. Pierre Mendès-France, IUT 2, Grenoble),
- Mustapha Rachdi (Univ. Pierre Mendès-France, Grenoble),

# Liste des conférenciers

- Samir BENAISSA (Univ. Sidi-Bel-Abbès)
- Delphine BLANKE (Univ. Paris 6)
- Hervé CARDOT (ENESAD-INRA, Dijon)
- Christophe CRAMBES (Univ. Paul Sabatier, Toulouse)
- Antonio CUEVAS (Univ. Autonome, Madrid)
- Pedro DELICADO (Univ. Polytechnique, Barcelone)
- Laurent DELSOL (Univ. Paul Sabatier, Toulouse)
- M'hammed EZZAHRIOUI (Univ. Littoral)
- Manuel FEBRERO (Univ. Santiago de Compostela)
- Arnaud GUYADER (Univ. Rennes 2)
- Alois KNEIP (Univ. Bonn)
- Abbes RABHI (Univ. Sidi-Bel-Abbès)
- Jérôme SARACCO (Univ. Dijon)
- Yingcai SU (Univ. Missouri State)
- Anne-Françoise YAO (Univ. Aix-Marseille 2).

| | Programme du Jeudi 15 Juin 2006 |
|---|---|
| 9h45-10h | Ouverture et Présentation des Journées |
| 10h00-10h30 | **Alois KNEIP**<br>Univ. de Bonn<br>*Common functional principal components* |
| | PAUSE CAFE |
| 10h45-11h15 | **Hervé CARDOT**<br>ENESAD, INRA Dijon<br>*Sur la prise en compte de covariables dans l'analyse des données fonctionnelles* |
| 11h15-11h45 | **Laurent DELSOL**<br>Univ. Toulouse 3<br>*Normalité asymptotique de la régression sur variable fonctionnelle avec application à la construction d'I.C.* |
| | PAUSE DEJEUNER |
| 13h45-14h15 | **Pedro DELICADO**<br>Univ. Politècnica de Catalunya<br>*Functional ANOVA when data are a weighted sample of density functions* |
| 14h15-14h45 | **Christophe CRAMBES**<br>Univ. Toulouse 3<br>*Smoothing splines estimators in functional linear regression with errors-in-variables* |
| 14h45-15h15 | **Samir BENAISSA**<br>Univ. Sidi Bel-Abbès<br>*Prédiction fonctionnelle pour les inclusions différentielles stochastiques* |
| | PAUSE CAFE |
| 15h30-16h00 | **Delphine BLANKE**<br>Univ. Paris 6<br>*Estimation de la régularité d'un processus Gaussien* |
| 16h00-16h30 | **M'hammed EZZAHRIOUI**<br>Univ. Littoral<br>*Normalité asymptotique du quantile conditionnel pour des données fonctionnelles dépendantes* |
| 16h30-17h00 | **Yingcai SU**<br>Univ. Missouri State<br>*Gene Expression Analysis between Vitis vinifera and the Disease-Resistant American Grapevine Vitis aestivalis* |

| | Programme du Vendredi 16 Juin 2006 |
|---|---|
| 9h00-9h30 | **Antonio CUEVAS**<br>Univ. Autonoma de Madrid<br>*On classification methods with functional data* |
| 9h30-10h00 | **Manuel FEBRERO**<br>Univ. de Santiago de Compostela<br>*Some ideas on depth for discrimination with functional data* |
| | $\mathfrak{P}$AUSE $\mathfrak{C}$AFE |
| 10h15-10h45 | **Jérome SARACCO**<br>Univ. de Bourgogne<br>*Une méthode SIR multivariée ($PMS_\alpha$)*<br>*en présence de covariables qualitatives* |
| 10h45-11h15 | **Anne Françoise YAO**<br>Univ. Aix-Marseille 2<br>*Modèle de régression inverse en*<br>*régression fonctionnelle* |
| 11h15-11h45 | **Arnaud GUYADER**<br>Univ. Rennes 2<br>*Classification par les plus proches voisins*<br>*en dimension infinie* |
| 11h45-12h15 | **Abbès RABHI**<br>Univ. Sidi Bel-Abbès (Algérie)<br>*Estimation of conditional distribution*<br>*and conditional hazard function* |
| 12h15-12h30 | $\mathfrak{F}$*ermeture des* $\mathfrak{Q}$*uatrièmes* $\mathfrak{J}$*ournées STAPH* |

10

.

# Prédiction fonctionnelle dans les inclusions différentielles stochastiques : Théorie et application

**Samir Benaissa*, Abdelghani Ouahab, Abderrahmane Yousfate**

* Adresse pour correspondance :
Laboratoire de Mathématique, BP 89
Université Djilali Liabès
Sidi Bel Abbès, 22000, Algérie.
e-mail : benaissa@univ-sba.dz , ouahab@univ-sba.dz , yousfate@univ-sba.dz

## Résumé

Les équations différentielles stochastiques à retard, appelées aussi équations différentielles fonctionnelles sont des équations du type

$$dX(t) = a(t, X(t))dt + \sigma(t, X(t))dw(t); \quad t \in J = [0, b] \tag{1}$$

avec la condition initiale

$$X(s) = \phi(s); \quad s \in ]-r, 0]$$

($r$ pouvant être infini) où $X$ et $a$ sont à valeurs dans un espace de Hilbert séparable H, $w$ est un processus de Wiener à valeurs dans un espace de Hilbert séparable $K$ et $\sigma$ est un opérateur linéaire continue de $K$ dans $H$. Notons $Q$ l'opérateur de covariance de $w$ dont nous considérons d'abord que le spectre est minoré par un scalaire $\varepsilon > 0$.

Le produit scalaire dans $K$ est défini comme suit $\forall k_1,\, k_2 \in K,\ \langle k_1, k_2\rangle_K = \langle Qk_1, k_2\rangle$, qui vérifie $E\langle w(t), k_1\rangle\langle w(s), k_2\rangle = (t \wedge s)\langle Qk_1, k_2\rangle$

Le produit scalaire dans $H$ est noté $\langle\ ,\ \rangle_H$ qui véfifie

$$\forall k_1,\, k_2 \in K,\ \langle \sigma k_1, \sigma k_2\rangle_H = \langle \sigma k_1, \sigma k_2\rangle = \langle \sigma^* \sigma k_1, k_2\rangle = \langle Q^{-1}\sigma^*\sigma k_1, k_2\rangle_K$$

La famille $\mathcal{F}_t$ (bornée supérieurement par $\mathcal{F}$) sur laquelle $X$ est adapté est continue à droite et vérifie $\mathcal{F}_{t-t'} \subset \mathcal{F}_t$ pour tout $t' \leq r$.

Nous noterons $M_2([-r, b], H)$ la classe des processus stochastiques à valeurs dans $H$ $\mathcal{F}_t$-adaptés dont les coordonnées hilbertiennes ont un moment de second ordre

$$\|X\|_2 = \left( \int_{-r}^{b} E|X(t)|^2 \, dt \right)^{\frac{1}{2}} < \infty.$$

L'espace $M_2$ ainsi construit est un espace hilbertien modulo l'équivalence p.p..

Les cas scalaires ou vectoriel de dimension fini ont fait l'objet de plusieurs travaux (**?**), (**?**), ... Si $a$ dans l'équation (1) est déterminée uniquement par des parties de $H$ (nous pouvons considérer les boréliens de $H$), nous utilisons les inclusions différentielles. L'équation (1) devient

$$dX(t) \in A(t, X(t))dt + \sigma(t, X(t))dw(t); \quad t \in J = [0, b]$$

où $A$ est une fonction multivoque à valeurs dans les parties de $H$ avec la même condition initiale.

Dans le cas théorique, nous présentons des conditions assez faibles sur $A$ pour l'existence d'une solution unique (au sens trajectoriel) pour l'équation (1) et nous étendons le résultat au cas où $Q$ est semi définie positive.

Dans la pratique, nous n'avons pas besoin d'avoir une solution trajectorielle unique, mais plutôt d'une tendance trajectorielle. Pour cela, nous présentons une construction permettant de choisir certains sous-espaces $H$ sur lesquels nous pouvons évaluer la tendance de $X(t)$ sur l'intervalle $[0, b]$.

Pour le processus solution, nous devons avoir

$$\forall t \in T, \forall \tau \in [0, b], \; E\left(X(t+\tau)/(X(s))_{s \leq t}\right) = E\left(X(t+\tau)/(X(s))_{s \in ]t-r, t]}\right) \quad (2)$$

que nous montrons que c'est un processus de Markov à retard.

Nous montrons également qu'en cas d'existence d'une solution $a \in A$, elle peut être évaluée avec des techniques non paramétriques.

# Références


M. Benchohra S.K. Ntouyas and A.Ouahab (2006). *On a nondensely defined semilinear stochastic functional differential equations with nonlocal conditions*, *J. Appl. Mathematics and Stochastic Analysis*, to appear

G. Da Prato and J. Zabczyk, (1992). *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge

I.I. Gikhman and A. Skorokhod, (1972). *Stochastic Differential Equations*, Springer-Velag.

A.A. Gushchin, U. Kücler, (1999). *Asymptotic inference for a linear stochastic differential equation with time delay* Bernoulli 5, No. 6, 1059-1098.

J. K. Hale, (1987). *Retarded equations with infinite delay, in Functional and Approximation of Fixed Points*, Proceedings, Bonn, Lecture Notes in Mathematics, **1223**, Springer Berlin, 61-63.

Y.A. Kutoyants, T. Mourid, D. Bosq, (1992). *Estimation paramétrique d'un processus de diffusion avec retards* Ann. IHP 28, 96-106.

A. Yousfate (2002). *Sur l'estimation fonctionnelle des opérateurs de transition des processus U-markoviens* Journées de statistique fonctionnelle et opératorielle, 11-12 juin 2002, Toulouse.

.

# Estimation du nombre de dérivées d'un processus Gaussien

**Delphine Blanke*** **et Céline Vial**

* Adresse pour correspondance :
Université Pierre et Marie Curie-Paris 6
L.S.T.A., 175 rue du Chevaleret, 8ème étage, Bat. A, 75013 Paris
e-mail : dblanke@ccr.jussieu.fr et cvial@u-paris10.fr

## Résumé

Soit $X = \{X_t,\, t \in [0,1]\}$ un processus Gaussien réel dont la dérivée en moyenne quadratique d'ordre $r_0$ ($r_0 \in \mathbb{N}$) satisfait une condition de type Hölder d'exposant $b_0 \in [0,1[$ ($r_0$, $b_0$ tous deux inconnus). On propose et on étudie un estimateur de $r_0$ basé sur les observations de $X$ aux instants $T = \{t_1, \ldots, t_N\} \subset [0,1]$.

Cette question ne semble pas avoir été réellement étudiée jusqu'à présent. Néanmoins, ce type de problème peut se poser naturellement dans des problèmes d'estimation ou bien de prévision où les processus sont supposés appartenir à des classes de régularités dépendant de $r_0$. Plus précisément, on peut se référer aux travaux de Cuzick (1977), Lindgren (1979), Bucklew (185) et Ditvelsen and Sorensen (2004) où les processus sont supposés avoir des dérivées en moyenne quadratique d'un ordre spécifique.

Dans une première partie, nous donnons les hypothèses générales portant sur $X$. Ces hypothèses seront en particulier satisfaites sous des conditions de type Sacks et Ylvisaker (SY) d'ordre $r_0$. Nous présentons alors des résultats sur l'interpolation polynomiale de Lagrange par morceaux de la trajectoire $X(t)$. Dans ce contexte, de nombreuses méthodes ont déjà été données et étudiées. Par exemple, sous les conditions de type SY, on peut citer les travaux de Müller-Gronbach (1996) (interpolation par projection orthogonale, plan d'échantillonnage optimal), Müller-Gronbach *et* al. (1997) (interpolation linéaire, plan d'échantillonnage optimal), Müller-Gronbach *et* al. (1998) (interpolation linéaire, plan d'échantillonnage adaptatif). Sous des conditions de type höldériennes, on peut également citer les articles de Seleznjev (1996) and (2000) traitant de l'interpolation linéaire (ou bien par polynômes d'Hermite).

Dans une deuxième partie, nous présentons une extension de ces derniers résultats en considérant des polynômes d'interpolation par morceaux $\widetilde{X}_r(t)$ de degré arbitraire $r \geq 1$. On montre essentiellement que l'erreur quadratique d'in-

terpolation décroît quand $r$ augmente mais se stabilise dès que $r \geq r_0$. Ce dernier point nous permet d'estimer $r_0$ par $\widehat{r}$ grâce à un critère empirique d'interpolation. Nous établissons, via une inégalité exponentielle pour $P(\widehat{r} \neq r_0)$, que $\widehat{r}$ converge presque sûrement vers $r_0$. Finalement, les résultats obtenus permettent également d'obtenir la convergence presque sûre de $\widetilde{X}_{\widehat{r}}(t)$ vers $X(t)$ avec une vitesse quasi-optimale.

# Résumé


Blanke, D. and Vial, C. (2006). Assessing the number of mean-square derivatives of a Gaussian process. Prepint.

Bucklew, J.A. (1985). A note on the prediction error for small time lags into the future. *IEEE Trans. Inform. Theory*, **31**(5), 677-679.

Cuzick, J. (1977). A lower bound for the prediction error of stationary Gaussian processes. *Indiana Univ. Math. J.*, **26**, 577-584.

Ditlevsen, S. and Sorensen, M. (2004). Inference for observations of integrated diffusion processes. *Scand. J. Statist.*, **31**(3), 417-429.

Lindgren, G. (1979). Prediction of level crossings for normal processes containing deterministic components. *Adv. in Appl. Probab.*, **11**(1), 93-117.

Müller-Gronbach, T. (1996). Optimal designs for approximating the path of a stochastic process. *J. Statist. Plann. Inference*, **49**(3), 371-385.

Müller-Gronbach, T. and Ritter, K. (1997). Uniform reconstruction of Gaussian processes. *Stochastic Process. Appl.*, **69**(1), 55-70.

Müller-Gronbach, T. and Ritter, K. (1998). Spatial adaption for predicting random functions *Ann. Statist.*, **26**(6), 2264-2288.

Seleznjev, O. (1996). Large deviations in the piecewise linear approximation of Gaussian processes with stationary increments. *Adv. in Appl. Probab.*, **28**(2), 481-499.

Seleznjev, O. (2000). Spline approximation of random processes and design problems. *J. Statist. Plann. Inference*, **84**(1-2), 248-262.

# Conditional functional principal component analysis

**Hervé Cardot**

* Adresse pour correspondance :
CESAER, UMR INRA-ENESAD,
26, bd Docteur Petitjean,
BP 87999, 21079 Dijon Cedex - France
e-mail : herve.cardot@enesad.inra.fr

## Résumé

Since the pioneer work by Deville (1974) much attention has been given to functional data analysis in the statistical community (see *e.g* Ramsay and Silverman 2002, 2005 and references therein). Many publications are devoted to the statistical description of a sample of curves (growth curves, temperature curves, spectrometric curves, ...) by means of the functional principal components analysis (Besse and Ramsay 1986, Castro *et al.* 1986, Kirkpatrick and Heckman, 1989, Rice and Silverman 1991, Kneip and Utikal 2001, ...). Performing the spectral decomposition of the empirical covariance operator, which is the analogous of the covariance matrix in a function space, allows to get a low dimensional space which exhibits, in a optimal way according to a variance criterion, the main modes of variation of the data. Let us consider a random function $Y(t)$ where index $t$ varies in a compact interval $T$ of $\mathbb{R}$, with mean $\mu(t) = \mathbb{E}(Y(t))$ and covariance function $\gamma(s,t) = \mathrm{Cov}(Y(s), Y(t)),\ s \in T$. Under general conditions (see *e.g.* Loève, 1978), the covariance function may be expressed as follows

$$\gamma(s,t) \;=\; \sum_{j\geq 1} \lambda_j \; v_j(s) \; v_j(t) \; , \tag{3}$$

where the $\lambda_j$ are the ordered eigenvalues, $\lambda_1 \geq \lambda_2 \geq ... \geq 0$, of the covariance operator and the functions $v_j$ the associated orthonormal eigenfunctions. Then, the best linear approximation $\widetilde{Y}^q$ to $Y$ in a function space with finite dimension $q$ is given by projecting the centered random function $Y - \mu$ onto the space generated by $\{v_1, \ldots, v_q\}$

$$\widetilde{Y}^q(t) \;=\; \mu(t) + \sum_{j=1}^{q} c_j \; v_j(t) \; . \tag{4}$$

where the random coordinates $c_j = \int_T (Y(t) - \mu(t))v_j(t)dt$, also called principal components (Dauxois *et al.*, 1982), are centered with variance $\text{var}(c_j) = \lambda_j$. This expansion is also known as the Karhunen-Loève expansion of $Y$ truncated at order $q$. The reader is referred to Loève (1978), Kirkpatrick & Heckman (1989) or Chiou *et al.* (2003b) for a comprehensive introduction on this topic.

This work aims at deriving a Karhunen-Loève expansion or FPCA which is able to take into account non-parametrically the effect of a quantitative covariate $X$ on $Y$ in order to get a decomposition similar to (4) that incorporates this additional information. Conditional on $X = x$, we would like to get the following optimal decomposition

$$\widetilde{Y}^q(x,t) \quad = \quad \mu(x,t) + \sum_{j=1}^{q} c_j(x)\ v_j(x,t)\ , \tag{5}$$

allowing the mean function and the basis functions $v_j(x,t)$ to depend non-parametrically on the covariate effect $x$.

The introduction of an additional information in such a framework has not received much attention in the literature whereas it can be very interesting in many situations. Silverman (1995) suggested a practical approach that could handle this kind of problem with parametric models. The estimation procedure is rather heavy and parametric models are not always adapted when one does not know in advance what can be the relationship between the dependent functional observations and the covariates. More recently, Chiou *et al* (2003b) considered a general approach that incorporates a covariate effect through a semi-parametric model. The problem was to estimate the number of eggs laid per day by $n = 936$ female Mediterranean fruit flies (see Carey *et al.* 1998 for a description of the experiments and of the data) for a time period restricted to the first 50 days of egg laying, conditional on the covariate $X$ which is the total number of eggs laid during the period. The mean function, that is to say the number of laid eggs per day during the first 50 days of lifetime, and the Karhunen-Loève basis are estimated on the whole population but the coordinates, *i.e.* the principal components, of an egg laying curve in this basis are obtained thanks to a single index model which take into account the covariate effect. A sample of 80 egg laying curves is drawn in Figure (1,(a)), showing a large variability in their shapes.

Instead of incorporating directly the covariate effect in the Karhunen-Loève expansion, we consider kernel regression estimators of the conditional expectation, $\mu(x) = \mathbb{E}(Y|X = x)$, and the conditional covariance operator,

$$\Gamma_x = \mathbb{E}((Y - \mu(x)) \otimes (Y - \mu(x))|X = x).$$

Then, we can derive estimators of the conditional eigenvalues $\lambda_j(x)$ and conditional eigenfunctions $v_j(x,t)$ by means of a spectral decomposition of $\Gamma_x$.

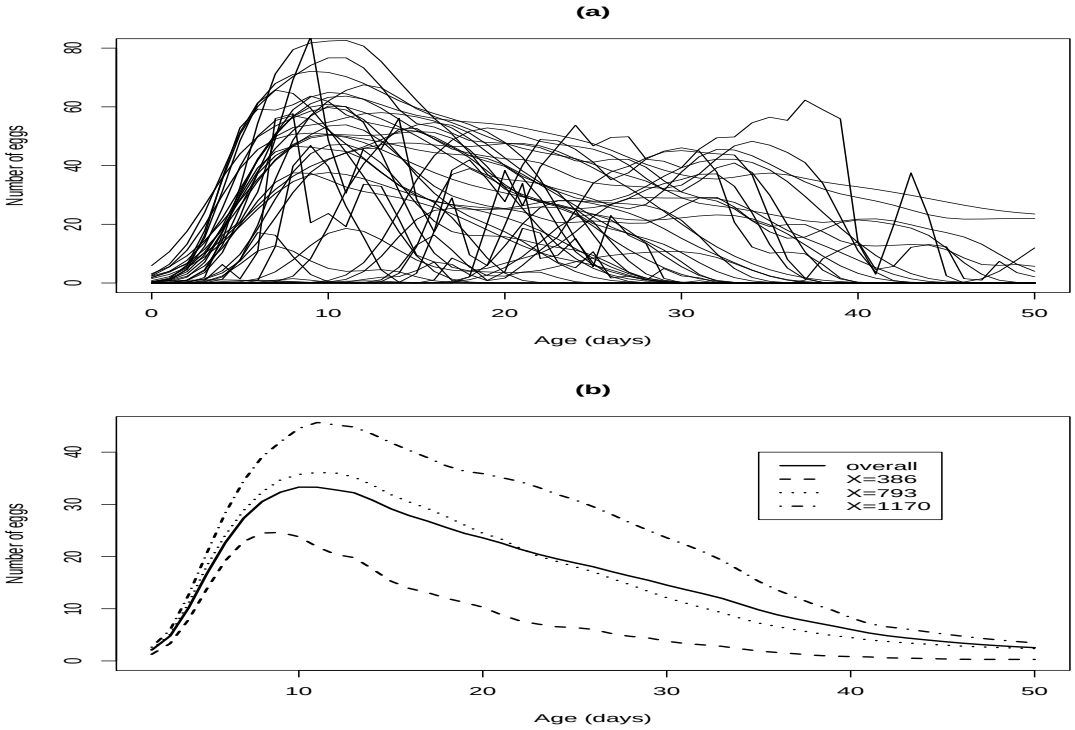


FIG. 1 – (a) : a sample of 80 smoothed egg laying curves. (b) : A comparison of the overall mean egg laying curve with conditional mean egg laying curves estimated for the first ($X = 386$), second ($X = 793$) and third ($X = 1170$) quartiles of the total number of eggs. Bandwidths values are selected by minimizing a cross-validation criterion.

# Conditional functional principal components analysis of the egg laying curves

We present now an illustration of the FPCA on egg laying curves data. The data consist of $n = 936$ egg laying curves of mediterranean fruit flies observed daily during the first 50 days of egg laying. The issue of determining how reproductive patterns are associated with overall reproductive success, measured by the total number of laid eggs during the period, is of real interest (Chiou *et al.*, 2003b).

The original curves are rather rough and a pre-smoothing step was performed by kernel regression using a Gaussian kernel. As noticed by Kneip and Utikal (2001) or Chiou *et al.* (2003b) under-smoothing seems to lead to better estimation in this framework and we consider, for each curve, three smoothed functional approximations based on individual smoothing parameters $h_{i,cv}$ chosen by minimizing a classical cross-validation criterion as well as under-smoothed approximations taking the bandwidths $h_{i,cv}/2$, $h_{i,cv}/3$ and $h_{i,cv}/6$.

The covariate $X$ which represents the total number of eggs has been normalized, without loss of generality, in order to take values in the interval $[0, 1]$. Before normalization, the mean number of eggs laid was 801, the first quartile was 386, the median was 793 and the third quartile 1170. The minimum value was 2 and the maximum was 2349.

The estimated conditional mean egg laying curves are drawn in Figure (1, (b)) for the first quartile, the median and the third quartile of the total number of laid eggs. It can be seen that their shapes clearly depend on the total number of eggs and that is why Chiou *et al.* (2003a) proposed a model on the mean function based on a multiplicative effect which seems to be adapted to that problem.

We can note that even if there are no large differences of cross-validated mean square errors between pre-smoothing approaches, functional approximations to the discretized curves obtained with small bandwidth values lead to better predictions and no smoothing should be preferred to usual smoothing procedures based on cross-validation. It appears that the most important tuning parameter is the parameter $h_1$ which controls the dependence of the mean function on the covariate $X$. Taking $h_1$ around 0.004 leads to a prediction error less than 228 for "under-smoothed" curves. If we consider the unconditional mean function, the leave out one curve criterion gives a prediction error around 352. Let us also remark that our functional approach performs well compared to those proposed by Chiou *et al.* (2003a, 2003b) whose best prediction error is around 315, according to the same criterion.

We also remark that a pre-smoothing step performed with small bandwidth values lead to better prediction but, as before, the most important tuning parameter seems to be $h_2$ which controls the effect of the covariate $X$ on the covariance function. The first and second conditional eigenfunctions are drawn in Figure (2, (c) and (d)) for three different values (first quartile, median and third quartile) of the total number of eggs and individual pre-smoothing steps performed with $h_{i,cv}/3$.

A comparison with the overall eigenfunctions clearly indicates a kind of "lag" in the dimension, meaning that the first eigenfunction try to capture the vertical shift information brought by the covariate which is already included in the conditional mean function. On the other hand, the second overall eigenfunction has roughly the same shape than the first conditional eigenfunctions. If we compare now the conditional eigenfunctions, it clearly appear that their shapes are different, and they take larger values, for fixed ages greater than 30 days, as the covariate increases. This means that larger variations occur at the end of the time interval when the number of laid eggs is large.

More details on estimation procedures of conditional FPCA as well as some asymptotic properties are given in Cardot (2006).

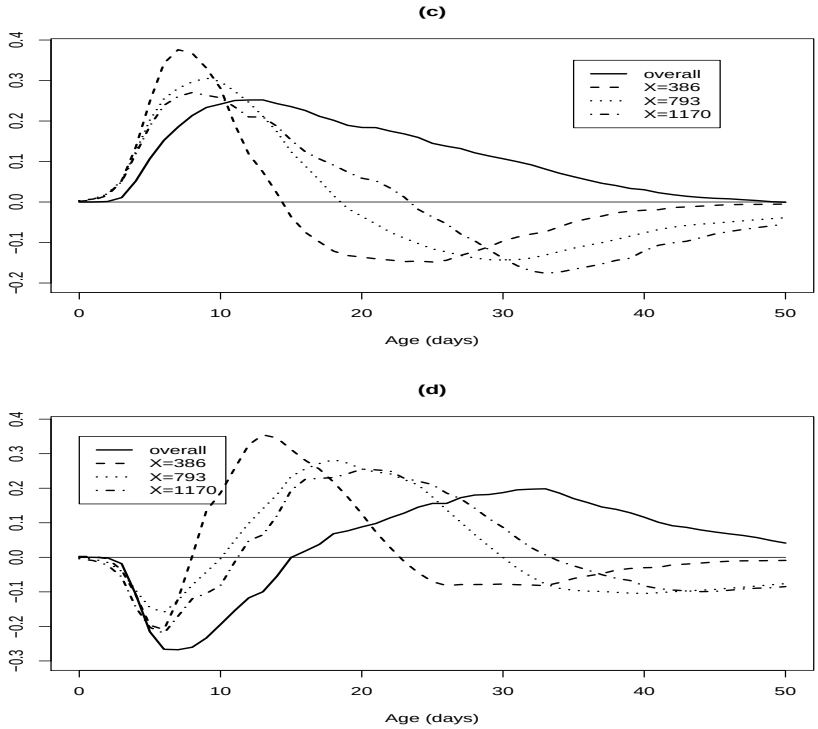


FIG. 2 – (a) : A comparison of the overall eigenfunctions and the conditional ones estimated for the first ($X = 386$), second ($X = 793$) and third ($X = 1170$) quartiles of the total number of eggs. (c) : First overall and conditional eigenfunctions. (d) : Second overall and conditional eigenfunctions. Bandwidths values are selected by minimizing a two steps cross-validation criterion.

# Références

Benko, M., Härdle, W. and Kneip, A. (2006). Common functional principal components. *SFB649 Economic Risk Discussion Paper 2006-10*, Humboldt University, Berlin.

Besse, P.C and Ramsay, J.O. (1986). Principal component analysis of sampled curves. *Psychometrika*, **51**, 285-311.

Cardot, H. (2000). Nonparametric estimation of the smoothed principal components analysis of sampled noisy functions. *Nonparametric Statistics*, **12**, 503-538.

Cardot, H., Faivre, R. and Maisongrande, P. (2004). Random Effects Varying Time Regression Models : Application to Remote Sensing. *Compstat 2004 proceedings, ed. J. Antoch, Physica Verlag,* 777-784.

Cardot, H. (2006). Conditional Functional Principal Components Analysis. *Scand. J. of Statistics*, à paraitre.

Carey, J.R., Liedo, P., Müller, H.G., Wang, J.L., Chiou, J.M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *J. of Gerontology Biological Sciences* **53**, 245-251.

Castro, P., Lawton, W. and Sylvestre, E. (1986). Principal Modes of Variation for Processes with Continuous Sample Curves. *Technometrics*, **28**, 329-337.

Chiou, J.M., Müller, H.G., Wang, J.L., Carey, J.R. (2003a). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. *Statistica Sinica* **13**, 1119-1133.

Chiou, J-M., Müller, H.G. and Wang, J.L. (2003b). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, B*, **65**, 405-423.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function : some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136-154.

Deville, J.C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee*, **15**, 3-104.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms and other infinite dimensional characters. *J. of Mathematical Biology*, **27**, 429-450

Kneip, A. and Utikal, K.J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *J. Amer. Statist. Assoc.*, **96**, 519-542.

Lecoutre, J.P. (1990). Uniform consistency of a class of regression function estimators for Banach-space valued random variable. *Statistics & Probability Letters*, **10**, 145-149.

Loève, M. (1978). *Probability Theory*, Springer, New-York.

Ramsay, J. O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer-Verlag, 2nd ed.

Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis : Methods and Case Studies.* Springer-Verlag.

Rice, J. and Silverman, B.W. (1991). Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves. *Journal of the Royal Statistical Society, B*, **53**, 233-243.

Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Silverman, B.W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, B*, **57**, 673-689.

Staniswalis, J.G., Lee, J.J. (1998). Nonparametric Regression Analysis of Longitudinal Data. *J. Amer. Statist. Assoc.*, **93**, 1403-1418.

# Smoothing Splines Estimators in Functional Linear Regression with Errors-in-Variables

**Hervé Cardot, Christophe Crambes *,**
**Alois Kneip and Pascal Sarda**

* Adresse pour correspondance :
Université Paul Sabatier, Laboratoire de Statistique et Probabilités, UMR C5583, 118, route de Narbonne, 31062 Toulouse Cedex, France
e-mail : crambes@cict.fr

## Introduction

In many applications (climatology, teledetection, linguistics, ...), data come from the observation of continuous phenomenons of time or space. These data are called *functional data* in the literature (for an overview on functional data analysis, see for instance Ramsay and Silverman, 2005). Many models for functional data have been studied, as for example the so-called *functional linear model* (see Ramsay and Dalzell, 1991), which will be the framework of this study. More precisely, we want to explain the effects of a variable $X$ on a variable $Y$, with $X$ taking its values in $L^2([0,1])$ (space of functions $f : [0,1] \longrightarrow \mathbb{R}$ such that $\int_0^1 f(t)^2 dt$ is finite), and $Y$ taking its values in $\mathbb{R}$. Then, we write

$$Y = \int_0^1 \alpha(t)X(t)dt + \epsilon, \tag{6}$$

where the function $\alpha \in L^2([0,1])$ is unknown and $\epsilon$ is a real random variable such that $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) = \sigma_\epsilon^2$. Considering data $(X_i, Y_i)_{i=1,\ldots,n}$, the goal is then to estimate the function $\alpha$. Let us notice here that the curves $X_i$ can be random or not. Situations where the $X_i$ are non random can be found, among others, in chemometrics where the $X_i$ correspond to responses obtained under predetermined experimental conditions (we find for example this situation in Cuevas et al., 2002). When the $X_i$ are random, we will consider a sample $X_1, \ldots, X_n$ of independent identically distributed variables. In any case, $Y_1, \ldots, Y_n$ are independent, and the expectations will always refer to the probability distribution induced by the random variable $\epsilon$. In the case where the $X_i$ are random, the expectations will be interpreted as conditional expectations given $X_1, \ldots, X_n$.

However, in the model (6), the variable $X$ is assumed to be observed without error, which may not be very realistic in practice, where many errors can prevent to know $X$ exactly. That is why we will consider that $X$ is not directly observed, but in fact we observe the variable $W = X+\delta$, where $\delta$ is a noise random variable. In practice, the curves are observed in some points $t_1 < \ldots < t_p$ of $[0,1]$. To get things simpler, we assume that these points are equally spaced, and the model we consider then writes

$$Y = \frac{1}{p}\sum_{j=1}^{p} \alpha(t_j)X(t_j) + \epsilon, \tag{7}$$

and, for $j = 1, \ldots, p$

$$W(t_j) = X(t_j) + \delta(t_j), \tag{8}$$

where $(\delta(t_j))_{j=1,\ldots,p}$ is a sequence of independent real random variables such that, for all $j = 1, \ldots, p$, $\mathbb{E}(\delta(t_j)) = 0$ and $\mathbb{E}(\delta(t_j)^2) = \sigma_\delta^2$. Equation (7) is an approximation of equation (6), the usual inner product of $L^2([0,1])$ being replaced by its discretized version. This approximation is of course valid for $p$ large enough, what we will assume from now on. In this framework, the goal is still to estimate the function $\alpha$, but using now the data $(W_i, Y_i)_{i=1,\ldots,n}$.

# Estimation of $\alpha$ in the non noisy case

Let us adopt the following notations : $\mathbf{Y} = (Y_1, \ldots, Y_n)^\tau$, $\boldsymbol{\alpha} = (\alpha(t_1), \ldots, \alpha(t_p))^\tau$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\tau$, where $\epsilon_i = Y_i - \int_0^1 \alpha(t)X_i(t)dt$ for all $i = 1, \ldots, n$. Moreover, we denote $\mathbf{X}$ the $n \times p$ matrix with general term $X_i(t_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Hence, the model writes

$$\mathbf{Y} = \frac{1}{p}\mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{9}$$

Many works already exist concerning the estimation of $\alpha$ (see for instance Cardot et al, 1999 and 2003). We choose here to build another estimator, based on *smoothing splines* (for an overview on smoothing splines, we refer to Eubank, 1988). We consider the space of *natural splines* of order $2m$ (with $m \in \mathbb{N}$) with knots in $t_1, \ldots, t_p$. It is a $p$ dimensional vectorial space, there exist basis functions $b_1, \ldots, b_p$. If we denote $\mathbf{b}(t) = (b_1(t), \ldots, b_p(t))^\tau$ and $\mathbf{B}$ the $p \times p$ matrix with general term $b_i(t_j)$ for $i, j = 1, \ldots, p$, we are looking for an estimator $\widehat{\boldsymbol{\alpha}}^*_{FLS,X}$ of $\boldsymbol{\alpha}$ solution of the minimization problem

$$\min_{\boldsymbol{a}\in\mathbb{R}^p}\left\{\frac{1}{n}\left\|\mathbf{Y}-\frac{1}{p}\mathbf{X}\boldsymbol{a}\right\|^2+\rho\boldsymbol{a}^\tau\mathbf{A}_m^*\boldsymbol{a}\right\}, \tag{10}$$

where $\mathbf{A}_m^* = \mathbf{B}\left(\mathbf{B}^\tau\mathbf{B}\right)^{-1}\left[\int_0^1 \mathbf{b}^{(m)}(t)\mathbf{b}^{(m)}(t)^\tau dt\right]\left(\mathbf{B}^\tau\mathbf{B}\right)^{-1}\mathbf{B}^\tau$ and $\|.\|$ is the usual euclidean norm of $\mathbb{R}^n$. This minimization problem has an explicit solution

$$\widehat{\boldsymbol{\alpha}}_{FLS,X}^* = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{X}^\tau\mathbf{X}+\rho\mathbf{A}_m^*\right)^{-1}\mathbf{X}^\tau\mathbf{Y}.$$

However, a problem appears because of the eigenvalues of the matrix $p\mathbf{A}_m^*$. This matrix has $m$ zero eigenvalues $\mu_{1,p}=\ldots=\mu_{m,p}=0$. Then, the existence of $\widehat{\boldsymbol{\alpha}}_{FLS,X}^*$ is not guaranteed. Nevertheless, if we consider $E_m$ the eigenspace corresponding to the $m$ zero eigenvalues $\mu_{1,p},\ldots,\mu_{m,p}$ and $\mathbf{P}_m$ the projection matrix on $E_m$, we set $\mathbf{A}_m=\mathbf{P}_m+p\mathbf{A}_m^*$ and our final estimator of $\boldsymbol{\alpha}$ will be

$$\widehat{\boldsymbol{\alpha}}_{FLS,X} = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{X}^\tau\mathbf{X}+\frac{\rho}{p}\mathbf{A}_m\right)^{-1}\mathbf{X}^\tau\mathbf{Y}, \tag{11}$$

solution of the minimization problem

$$\min_{\boldsymbol{a}\in\mathbb{R}^p}\left\{\frac{1}{n}\left\|\mathbf{Y}-\frac{1}{p}\mathbf{X}\boldsymbol{a}\right\|^2+\frac{\rho}{p}\boldsymbol{a}^\tau\mathbf{A}_m\boldsymbol{a}\right\}.$$

We give an asymptotic result for this estimator. This result is given for the semi-norm defined by $\|\boldsymbol{u}\|_\Gamma^2=\frac{1}{p}\boldsymbol{u}^\tau\left(\frac{1}{np}\mathbf{X}^\tau\mathbf{X}\right)\boldsymbol{u}$ for all $\boldsymbol{u}\in\mathbb{R}^p$. We will need the following assumption.
$(H.1)$ $\alpha$ is $m$ times differentiable and $\alpha^{(m)}$ belongs to $L^2([0,1])$.

**Theorem 1** *Under hypothesis* $(H.1)$*, there exists constants* $C_1, C_2 > 0$ *such that*

$$\|\mathbb{E}(\widehat{\boldsymbol{\alpha}}_{FLS,X})-\boldsymbol{\alpha}\|_\Gamma^2 \le \rho C_1, \tag{12}$$

*and*

$$\frac{1}{p}\mathbb{E}\left(\|\widehat{\boldsymbol{\alpha}}_{FLS,X}-\mathbb{E}(\widehat{\boldsymbol{\alpha}}_{FLS,X})\|^2\right) \le \frac{\sigma_\epsilon^2}{n\rho}C_2. \tag{13}$$

The proof of this result is given in Cardot et al. (2006).

# Estimation of $\alpha$ in the noisy case

Using the previous matricial notations, if in addition $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))^\tau$ and $\mathbf{W}_i = (W_i(t_1), \ldots, W_i(t_p))^\tau$ for all $i = 1, \ldots, n$, let $\mathbf{W}$ and $\boldsymbol{\delta}$ be the $n \times p$ matrices with respective general terms $W_i(t_j)$ and $\delta_i(t_j)$ (with $\delta_i(t_j) = W_i(t_j) - X_i(t_j)$ for all $i = 1, \ldots, n$ and for all $j = 1, \ldots, p$). The model is now

$$\left\{ \begin{array}{l} \mathbf{Y} = \frac{1}{p}\mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \\ \mathbf{W} = \mathbf{X} + \boldsymbol{\delta}, \end{array} \right.$$

what leads us to consider the minimization problem

$$\min_{\boldsymbol{a}\in\mathbb{R}^p, \mathbf{X}_i \in \mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[ \left( Y_i - \frac{1}{p}\mathbf{X}_i^\tau \boldsymbol{a} \right)^2 + \frac{1}{p} \left\| \mathbf{X}_i - \mathbf{W}_i \right\|^2 \right] + \frac{\rho}{p}\boldsymbol{a}^\tau \mathbf{A}_m \boldsymbol{a} \right\}. \tag{14}$$

We adapt here the so-called *total least squares* method already studied in a multivariate context (see for example the works in Golub and Van Loan, 1980 or Van Huffel and Vandewalle, 1991). Then, we have the following result (with the proof in Cardot et al, 2006).

**Proposition 0.0.1** *The solution of the minimization problem (14) is given by*

$$\widehat{\boldsymbol{\alpha}}_{FTLS} = \frac{1}{np}\left( \frac{1}{np^2}\mathbf{W}^\tau\mathbf{W} + \frac{\rho}{p}\mathbf{A}_m - \sigma_k^2 \mathbf{I}_p \right)^{-1} \mathbf{W}^\tau \mathbf{Y}, \tag{15}$$

*where* $\mathbf{I}_p$ *is the identity matrix of size* $p$ *and* $\sigma_k^2$ *is the smallest non-zero eigenvalue of the matrix*

$$\frac{1}{n}\left( \frac{\mathbf{W}}{p}, \mathbf{Y} \right)^\tau \left( \frac{\mathbf{W}}{p}, \mathbf{Y} \right) + \frac{\rho}{p} \left( \begin{array}{cc} \mathbf{A}_m & \mathbf{0} \\ \mathbf{0} & 0 \end{array} \right).$$

As numerical problems can appear (due to the fact that the eigenvalues of this matrix can decrease rapidly to zero), we rather choose to consider the estimator

$$\widehat{\boldsymbol{\alpha}}_{FTLS} = \frac{1}{np}\left( \frac{1}{np^2}\mathbf{W}^\tau\mathbf{W} + \frac{\rho}{p}\mathbf{A}_m - \frac{\widehat{\sigma}_\delta^2}{p^2} \mathbf{I}_p \right)^{-1} \mathbf{W}^\tau \mathbf{Y}, \tag{16}$$

where $\widehat{\sigma}_\delta^2$ is an estimator of $\sigma_\delta^2$. We use here is a nonparametric estimator studied in (**?**) and given by

$$\widehat{\sigma}_\delta^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{6(p-2)} \sum_{j=2}^{p-1} \left[ W_i(t_{j-1}) - W_i(t_j) + W_i(t_{j+1}) - W_i(t_j) \right]^2. \tag{17}$$

Finally, we present an asymptotic result for the estimator $\widehat{\boldsymbol{\alpha}}_{FTLS}$. More precisely, we compare this estimator to the one presented in the previous section, when the curves $X_i$ are non noisy and then directly observable. We consider the following hypotheses.
$(H.2)$ $\sup_i \sup_j |X_i(t_j)| \leq C_3 < +\infty$ (or $P(\sup_i \sup_j |X_i(t_j)| \leq C_3) = 1$ in the case where the $X_i$ are random).
$(H.3)$ There exists a constant $C_4 > 0$ such that

$$\sup_{r,s=1,\ldots,p} \mathbb{E}\left(\delta_i(t_r)^2 \delta_i(t_s)^2\right) \leq C_4.$$

**Theorem 2** *Under hypotheses* $(H.1)$ - $(H.3)$*, if we also assume that* $Y_i \perp\!\!\!\perp \delta_i$ *for all* $i = 1, \ldots, n$ *and that* $\mathbb{E}(Y_i^2) < +\infty$*, then we have*

$$\|\widehat{\boldsymbol{\alpha}}_{FTLS} - \widehat{\boldsymbol{\alpha}}_{FLS,X}\| = O_P\left(\frac{\sigma_\delta}{n^{1/2} p^{1/2} \rho}\right). \tag{18}$$

The proof of this result can also be found in Cardot et al (2006).

# Références


Cardot, H., Crambes, C., Kneip, A. and Sarda, P. (2006). Smoothing Splines Estimators in Functional Linear Regression with Errors-in-Variables. Submitted to *Computational Statistics and Data Analysis*.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional Linear Model. *Statistics and Probability Letters*, **45**, 11-22.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica*, **13**, 571-591.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear Functional Regression : the case of a Fixed Design and Functional Response. *Canadian Journal of Statistics*, **30**, 285-300.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual Variance and Residual Pattern in Nonlinear regression. *Biometrika*, **3**, 625-633.

Golub, G.H. and Van Loan, C.F. (1980). An Analysis of the Total Least Squares Problem. *SIAM, Journal of Numerical Analysis*, **17**, 883-893.

Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for Functional Data Analysis. *Journal of the Royal Statistical Society, Series B*, **53**, 539-572.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer, New York.

Van Huffel, S. and Vandewalle, J. (1991). *The Total Least Squares Problem : Computational Aspects and Analysis*. SIAM, Philadelphia.

# On classification methods with infinite-dimensional data

**Antonio Cuevas *,**
**José Ramon Berrendero, Manuel Febrero,**
**Ricardo Fraiman and Alberto Rodriguez-Casal**

*Adresse pour correspondance :
Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid, 28049-Madrid (Spain)
e-mail : antonio.cuevas@uam.es

## Abstract

This communication is concerned with supervised classification methods with infinite-dimensional data (to be more specific, the data are either real-valued functions or sets in the Euclidean space). The terms "discrimination", "statistical learning" or "machine learning" are also used, depending on the context, as approximate synonyms for "supervised classification".

The books by Ramsay and Silverman (2005) and Ferraty and Vieu (2006) provide some insights on this topic. Some papers on the subject are Ferraty and Vieu (2003) and James and Hastie (2001).

We will consider two different situations, which are briefly outlined in the following sections. First, we propose a method, based on the use of projections, to classify images (identified with compact sets in the Euclidean space). Second, we discuss a procedure to classify functions based on different notions of functional depth.

## Discrimination for sets

Let us assume we have two "training samples" of compact sets, in $[-1,1]^d$, drawn, respectively, from two "populations of images" $P_1$, $P_2$ :

$$S_1, \ldots, S_n$$
$$T_1, \ldots, T_m$$

We wish to classify a new set $C$ in either $P_1$ or $P_2$. Every set $S_i$, $T_j$ is assumed to be a possible support of an absolutely continuous distribution.

The basic idea is to associate with every set a random variable $X$ whose distr ibution can be evaluated. This will give two sets of densities

$$f_1, \ldots, f_n$$
$$g_1, \ldots, g_m$$

Now, we can use the $f_i$ and the $g_j$ in the classification process.

A first proposal is using the **distance to the boundary** of $[-1,1]^d$ of a point $\xi$ randomly chosen on the sets $S_i, T_j$ to define the random variables $X_i, Y_j$ associated with the $S_i$, $T_j$. Note that the densities of these variables can be approximated with arbitrary precision by Monte Carlo sampling.

A further possibility is choosing a set of directions $a_1, \ldots, a_k$ and evaluating, by Monte Carlo sampling, the densities $f_{i1}, \ldots, f_{in}$ of the **projections** $a_i' X_l$, $i = 1, \ldots, k$ where $X_l$ is uniform on $S_l$. The corresponding densities $g_{i1}, \ldots g_{im}$ are defined in a similar way for the $T_j$. Some previous standardization could be needed in order to make the procedure affine invariant.

The classification process can be made along the following lines :

– Define the averages

$$f_i = \frac{1}{n} \sum_{j=1}^{n} f_{ij}.$$

The definition of $g_i$ is analogous using the densities $g_{ij}$.

– In order to classify the set $C$ take a sample $x_1, \ldots x_B$ uniform on $C$ and evaluate the "membership score" to each group (either $S$ or $T$) (along the direction $a_i$) by

$$D_i(C;1)\#\{j : f_i(x_j) > g_i(x_j)\}, \quad D_i(C;2)\#\{j : f_i(x_j) \leq g_i(x_j)\}.$$

Of course, if the prior probability that $C$ to both groups is different we should modify this criterion accordingly.

– Finally, the "total scores" of each group are

$$D(C;1) = \sum_i D_i(C;1), \; D(C;2) = \sum_i D_i(C;2)$$

and we assign $C$ to the group with higher score.

# Discrimination in FDA based on data depth

Let us assume we have a training sample $(X_i, Y_i)$, $i = 1, \ldots, n$ where $X_i$ are functional data and $Y_i \in \{0, 1\}$. Now, given a new coming observation $X$, we

want to predict the corresponding value of the classification variable $Y$. Denote by $G_0$ and $G_1$ the groups of observations with $Y = 0$ and $Y = 1$, respectively.

The **main idea** is to classify the datum $X$ in the group $G_0$ when its depth inside $G_0$ larger than its depth inside $G_1$.

We will do this by defining a new concept of data depth which takes also into account the information provided by the derivatives (see also the communication by Manuel Febrero for related ideas).

**Integrated simplicial depth**

Given a functional sample $x_1(t), \ldots, x_n(t)$, we consider

$$z_1(t) = (x_1(t), \dot{x}_1(t)), \ldots, z_n(t) = (x_n(t), \dot{x}_n(t))$$

and evaluate the simplicial depth, $D_n(x(t))$ of $z(t) = (x(t), \dot{x}(t))$.

Finally, define the depth of $x$ by

$$I_n(x) = \int_0^1 D_n(x(t))dt \tag{19}$$

This notion of functional depth would take also into account the information provided by the derivatives.

The simplicial depth $D_n(x)$ allows for a population version, defined as the probability that $x$ belongs to a triangle of vertices $z_i(t) = (x_i(t), \dot{x}_i(t))$, where $x_i$, $i = 1, 2, 3$ are iid observations from the original process. Thus there is also a population version for $I_n$.

**Functional data depth using random directions**

Given a process $X$ as well as an independent "direction process" $H$, let $F[\langle h, X\rangle](.)$ the distribution function of the projection of $X$ along the "direction" given by the trajectory $h$ of $H$.

Notation : $D[\langle h, X\rangle](t) = \min\{F[\langle h, X\rangle](t), 1 - F[\langle h, X\rangle](t^-)\}$ (we might also take

$$D[\langle h, X\rangle](t) = F[\langle h, X\rangle](t)(1 - F[\langle h, X\rangle](t^-))$$

as the respective maxima coincide).

We define the population deepest trajectory as any function in the set

$$\text{argmax}_\mu E\left(D[\langle H, X\rangle](\langle H, \mu\rangle)\right). \tag{20}$$

Some remarks :

– Note that if $X$ has an **angular symmetric distribution**, i.e., there exists a function $\mu = \mu(t)$ such that

$$\frac{X - \mu}{\|X - \mu\|} \quad \text{and} \quad \frac{\mu - X}{\|X - \mu\|}$$

are equally distributed, then for each fixed $h$, $\mu$ minimizes (in $m$) $D[\langle h, m\rangle]$ and therefore $\mu$ is the only element in the set of deepest points (20).
- In particular, this condition is fulfilled for any non-degenerate Gaussian process.
- The definition (20) makes also sense for the finite-dimensional case thus giving a new concept of multivariate depth.

The empirical version :
Let us define

$$D_n[y_1, \ldots, y_n](t) = \min\{F_n[y_1, \ldots, y_n](t), 1 - F_n[y_1, \ldots, y_n](t^-)\},$$

where $F_n[y_1, \ldots, y_n](t)$ is the empirical distribution function associated with the sample of real values $y_1, \ldots, y_n$.

Now, we define the sampling deepest function to be the trajectory $x_{k_0}$ maximizing in $k$ the Functional Random Projection Depth

$$\frac{1}{N}\sum_{j=1}^{N} D_n[\langle h_j, x_1\rangle, \ldots, \langle h_j, x_n\rangle](\langle h_j, x_k\rangle),$$

where $h_i, \ldots h_N$ is a random sample of trajectories of the process $H$.

Some remarks :
- The empirical version is defined by maximizing on the sample, rather than on the whole space. Under general conditions, this should converge to the population version (20).
- The definition of the empirical deepest function depends (at least in principle) on the projection trajectories $h_1, \ldots, h_N$. This is in the spirit of the resampling techniques. Even the population version depends on the "projection process" $H$. Recall, however that the population deepest point is defined in a unique way, not depending on $H$ for the case of angularly symmetric processes.
- This notion of depth makes sense for any random process taking values in a separable Hilbert space (not only in a Hilbert function space), for example in an space of covariance operators.

**Functional data depth using random directions when we consider a random process and its derivative**

When we want to consider depth involving the random process and its derivatives, the random projection method can also be adapted in a quite natural way as follows.
Given a functional sample $x_1(t), \ldots, x_n(t)$, and a direction $h$ (a trajectory of $H$), we define

$$z_{1h} = (\langle h, x_1\rangle), \langle h, \dot{x}_1\rangle), \ldots, z_{nh} = (\langle h, x_n\rangle), \langle h, \dot{x}_n\rangle))$$

and evaluate the simplicial depth, $D_{nh}(x)$ of $z_h = (\langle h, x\rangle), \langle h, \dot{x}\rangle)$.
Finally, define the random projection depth of $x$ by

$$I_n(x) = \frac{1}{N}\sum_{j=1}^{n} D_{nh_j}(x). \tag{21}$$

**Two-steps random projections method (2-FRPD) for functional data**

Another possibility is to replace the use of the simplicial depth for a further application of the projections method, which could be used (in a second step) for the bi-dimensional projections

$$z_{1h} = (\langle h, x_1\rangle), \langle h, \dot{x}_1\rangle), \ldots, z_{nh} = (\langle h, x_n\rangle), \langle h, \dot{x}_n\rangle)).$$

This seems to have at least two advantages :

– Computational simplicity
– It is possible to incorporate additional derivatives without a significant additional computational burden. This is interesting, for example, in the study of the growth data considered by Ramsay.

# Références

Ramsay, J.O. and Silverman, B.W. (1997). *Functional data analysis*. Springer.

Ferraty, F. and Vieu, P. (2003). Curves discrimination : a nonparametric functional approach. Computational Statistics and Data Analysis 44, 161-173.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer (to appear).

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society B 63, 533-550.

.

# Functional ANOVA when data are a weighted sample of density functions

**Pedro Delicado***

* Adresse pour correspondance :
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
Despatx 214, Edifici C5, Campus Nord
C/ Jordi Girona 1-3, 08034 Barcelona (SPAIN)
e-mail : pedro.delicado@upc.edu

## Abstract

This paper deals with the ANOVA test for functional data when the observations are density functions and they have different weights. Two known ANOVA tests for generic functional data are adapted to this particular kind of data. Moreover, we introduce new procedures based on distances between pairs of observed density functions, allowing us to use the $L_1$ distance, the most natural for density functions. A simulation study is carried out to compare the practical behaviour of the available tests. Theoretical derivations has been done in order to allow weighted samples in the tests procedures. The paper ends with a real data example : for a collection of European regions we estimate the regional relative income densities and then we test the significance of the *country effect*.

## Introduction

In the last years, the joint development of real-time measurement instruments and data storage computer resources has made possible to observe and save complete functions as results of random experiments. Ramsay and Silverman (1997) express it saying that random functions are the *statistical atoms* in these cases. Functional Data Analysis (FDA) deals with the statistical description and modelization of samples of random functions.

The one-way analysis of variance is one of the standard methods that have been generalized to be used in FDA, giving rise to *functional ANOVA*. It is

assumed that $n$ functions $f_{ri}(x)$ have been observed from the model

$$f_{ri}(x) = m_r(x) + e_{ri}(x), i = 1, \ldots, n_r, r = 1, \ldots, k, x \in [a, b]\,,$$

where $\sum_r n_r = n$ and $m_r(x)$ are $k$ unknown mean functions and $e_{ri}(x)$ are independent trajectories drawn from a process with zero mean and covariance function $Cov(e_{ri}(x), e_{ri}(y)) = K(x, y)$. The hypothesis to be tested is $H_0 : m_1(x) = \cdots = m_k(x)$, for all $x \in [a, b]$. Two known methods for functional ANOVA are the proposals of Ramsay and Silverman (1997) (based on F-ratio functions) and that of Cuevas, Febrero and Fraiman (2004) (based on a functional Central Limit Theorem).

# Functional ANOVA for weighted density functions

This work has two main objectives. The first one is to fit functional ANOVA to data being density functions. We analyze the applicability of the two known techniques, see Ramsay and Silverman (1997) and Cuevas, Febrero and Fraiman (2004), and how they can be adapted to the case of densities. Moreover we present a distance based ANOVA test working on pairwise distances between observed data (Gower and Krzanowski, 1999). This device allows us to use the $L_1$ distance, that is the most natural one for density functions. The same distance based ANOVA procedure is also applicable to other distance definitions and to data not being density functions. The second objective is to generalize functional ANOVA tests for weighted samples. We establish results valid for weighted data that generalize those of Gower and Krzanowski (1999) and Cuevas, Febrero and Fraiman (2004).

The null distribution of any of the test statistics included in the paper is unknown. Thus, the use of permutations is required (see Gower and Krzanowski, 1999, and references therein). In this paper we consider two alternatives that we call *permuting observations* (the standard one) and *permuting residuals* (the model is estimated assuming the alternative hypothesis is truth, and then the artificial samples are defined functions as the sum of the global mean plus the permuted estimated residuals).

A simulation study is carried out to compare the practical behaviour of these tests. Moreover, for a collection of European regions we estimate the regional relative income densities and then we test the significance of the *country effect*. An extended version of this work is available in Delicado (2005).

**Acknowledgement** Research supported by the Spanish Ministry of Education and Science and FEDER, MTM2005-02370, and by the EU PASCAL Network of Excellence, IST-2002-506778.

# Références

Cuevas, A., M. Febrero, and R. Fraiman (2004). An anova test for functional data. *Computational Statistics and Data Analysis 47*, 111–122.

Delicado, P. (2005). Functional ANOVA when data are a weighted sample of density functions.
<www-eio.upc.es/~delicado/my-public-files/FANOVA.density.pdf>

Gower, J. C. and W. J. Krzanowski (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society, Series B 48*, 505–519.

Ramsay, J. and B. W. Silverman (1997). *Functional Data Analysis*. New York : Springer.

.

# Normalité asymptotique de la régression sur variable fonctionnelle avec application à la construction d'intervalles de confiance

**Laurent Delsol**

* Adresse pour correspondance :
Université Paul Sabatier
31062 Toulouse, France
e-mail : delsol@cict.fr

## Introduction

Les progrès réalisés au niveau des appareils de mesure entrainent que les données dont dispose le statisticien sont de moins en moins discrétisées. Par conséquent une nouvelle branche des statistiques étudiant des données fonctionnelles a vue le jour. On étudiera plus précisément dans cet exposé un modèle de régression non-paramétrique fonctionnelle avec des données dépendantes. Cet exposé complète les travaux de Masry (2005) et Ferraty et al. (2006) au sujet de la normalité asymptotique de l'estimateur à noyau fonctionnel pour des données $\alpha$-mélangeantes. L'explicitation des constantes qui rentrent en compte dans l'expression des termes dominants du biais et de la variance de l'estimateur à noyau permet d'établir des intervalles de confiance ponctuels pour la fonction de régression.

## Le modèle

Nous allons étudier tout au long de cet exposé le modèle de régression suivant : $Y = r(X) + \epsilon$, où $Y$ est une variable aléatoire réelle et $X$ une variable aléatoire à valeurs dans l'espace semi-métrique fonctionnel $(E, d)$. On suppose également que l'erreur $\epsilon$ est non corrélée avec la variable explicative $X$ et qu'elle a un moment conditionnel d'ordre deux par rapport à celle-ci. On ne fait aucune hypothèse sur la forme de l'opérateur $r$, seulement des hypothèses de régularité, c'est pourquoi notre modèle est dit non paramétrique. Enfin, on considère un échantillon constitué de $n$ paires $(X_i, Y_i)$ qui suivent la même loi que $(X, Y)$ qui peuvent être dépendantes (nous verrons comment par la suite).

On considère un élément $x$ de l'espace $E$ et on estime $r(x)$ par l'estimateur à noyau suivant :

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(X_i,x)}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{d(X_i,x)}{h_n}\right)}.$$

Cet estimateur a été introduit par Ferraty et Vieu (**?**) comme une généralisation au cas fonctionnel de celui de Nadaraya-Watson.

# Résultats principaux

**Notations et principales hypothèses**

Pour commencer, il nous faut introduire quelques fonctions qui joueront un rôle clef dans nos résultats.

$$\phi(s) = \mathbb{E}\left[(r(X) - r(x))/d(X,x) = s\right],$$
$$F(h) = \mathbb{P}(d(X,x) \leq h),\ \tau_h(s) = \frac{F(hs)}{F(h)}.$$

Nos premières hypothèses portent sur le modèle :

$$r \text{ est borné sur un voisinage de } x, \tag{22}$$
$$\phi(0) = 0 \text{ et } \phi'(0) \text{ existe}, \tag{23}$$
$$\sigma_\epsilon^2(\,.\,) \text{ est continue sur un voisinage de } x \text{ et } \sigma_\epsilon^2 := \sigma_\epsilon^2(x) > 0, \tag{24}$$
$$\forall s \in [0;1],\ \lim_{n\to+\infty} \tau_{h_n}(s) = \tau_0(s) \text{ with } \tau_0(s) \neq 1_{[0;1]}(s). \tag{25}$$

L'hypothèse (23) permet de donner l'expression exacte des constantes là où une hypothèse standard de type Lipschitz ne donnerait que des ordres de grandeur. On pourra aussi remarquer que les hypothèses faites sur la loi de $X$ ne se font qu'à travers des probabilités de petites boules et que la condition (25) est vérifiée par de nombreux processus standards.
Ensuite, il nous faut faire des hypothèses pour pouvoir contrôler la somme des covariances :

$$\exists p > 2,\ \exists M > 0,\ \mathbb{E}\left[|\epsilon|^p / X\right] \leq M \ a.s., \tag{26}$$
$$max\left(\mathbb{E}\left[|Y_i Y_j| / X_i, X_j\right], \mathbb{E}\left[|Y_i| / X_i, X_j\right]\right) \leq M a.s.\ \forall i,j \in \mathbb{Z}. \tag{27}$$

Enfin il nous faut faire quelques hypothèses concernant notre estimateur à noyau :

$$h_n = O\left(\frac{1}{\sqrt{nF(h_n)}}\right),\ \lim_{n\to+\infty} nF(h_n) = +\infty, \tag{28}$$

$$K \text{ a un support compact, est } C^1 \text{ et décroissante sur } ]0;1[\ ,\ K(1) > 0 \tag{29}$$

Nos résultats seront exprimés à l'aide des constantes suivantes :

$$M_0 = \left( K(1) - \int_0^1 (sK(s))' \tau_0(s)\, ds \right) \quad , \quad M_1 = \left( K(1) - \int_0^1 K'(s) \tau_0(s)\, ds \right)$$
$$et \;\; M_2 \;\; = \;\; \left( K^2(1) - \int_0^1 \left(K^2\right)'(s) \tau_0(s)\, ds \right).$$

Ce qui diffère du cas indépendant étudié par Ferraty et al. (2006) est qu'il nous faut des hypothèses supplémentaires concernant la dépendance de nos variables. On suppose que les $n$ paires $(X_i, Y_i)$ sont $\alpha$-mélangentes au sens de Rosenblatt (1956) et avec les coefficients définis par Rio (1995, p34, Notation 2.1). Il nous faut encore introduire quelques notations :

$$\Theta(s) := \max\left( \max_{i \neq j} P\left(d(X_i, x) \leq s,\, d(X_j, x) \leq s\right), F^2(s) \right),$$
$$\Gamma_i := Y_i K\left(\frac{d(X_i, x)}{h_n}\right),\; \Delta_i := K\left(\frac{d(X_i, x)}{h_n}\right), U_{i,n} = \frac{\Gamma_i \mathbb{E}[\Delta_i] - \Delta_i \mathbb{E}[\Gamma_i]}{F(h_n)\sqrt{nF(h_n)}}.$$

S'ajoutent également des hypothèses exprimées tout d'abord de manière très générale :

$$\exists (u_n)_{n\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}},\; O\left( \frac{n\left[\alpha(u_n)\right]^{\frac{p-2}{p}}}{F(h_n)^{\frac{p-2}{p}}} \right) + O\left( u_n \frac{\Theta(h_n)}{F(h_n)} \right) \overset{n\to+\infty}{\to} 0, \tag{$H1$}$$

et

$$I_n := n \int_0^1 \alpha^{-1}\left(\frac{x}{2}\right) Q^2_{U_{1,n}}(x)\, inf\left( \frac{3M_1\sqrt{\sigma_\epsilon^2 M_2}}{2}, \alpha^{-1}\left(\frac{x}{2}\right) Q_{U_{1,n}}(x) \right) dx \to 0, \tag{$H2$}$$

où $Q_{U_{i,n}}(x) = inf\{t,\, \mathbb{P}(|U_{i,n}| > t) \geq x\}$, et $\alpha^{-1}\left(\frac{x}{2}\right) = inf\left\{t,\, \alpha([t]) \geq \frac{x}{2}\right\}$. qui seront spécifiées dans deux cas plus particuliers.

**Normalité asymptotique**

**Théorème 1** *Normalité asymptotique :*
*Sous les hypothèses (22)-(29), $(H1)$ et $(H2)$, on obtient*

$$\frac{M_1}{\sqrt{M_2\sigma_\epsilon^2}} \sqrt{n\hat{F}(h_n)}\left(\hat{r}(x) - r(x) - B_n\right) \to N(0,1)$$

*où* $B_n = h_n \phi'(0) \frac{M_0}{M_1}$ *et* $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^n 1_{[d(X_i,x),+\infty[}(t)$.

Remarque :

– On peut remarquer que dans le cas où les coefficients de mélange sont arithmétiques d'ordre $a : \alpha(i) \leq Ci^{-a}$, les conditions (H1) et (H2) peuvent être remplacées par les suivantes :

$$\exists \nu > 0,\ \Theta(h_n) = O\left(F(h_n)^{1+\nu}\right), \text{ avec } a > \frac{(1+\nu)p-2}{\nu(p-2)}, \tag{30}$$

$$\exists \gamma > 0,\ nF^{1+\gamma}(h_n) \to +\infty \text{ et } a > max\left(\frac{4}{\gamma}, \frac{p}{p-2} + \frac{2(p-1)}{\gamma(p-2)}\right). \tag{31}$$

– Puisque les coefficients géométriques,c'est à dire tels que $\alpha(n) \leq C\rho^n$, sont arithmétiques d'ordre $a$ quelque soit $a$, on obtient le même résultat sans condition sur $\rho$ pour ce type de mélange.

**Intervalles de confiance asymptotiques**

Pour obtenir des intervalles de confiance asymptotiques, il nous faut estimer les constantes que nous avons introduites. S'il paraît assez simple d'estimer $M_1$ et $M_2$, le terme de biais pose plus de problèmes car on doit l'estimer avec une vitesse de convergence assez importante pour que l'on puisse lui faire remplacer $B_n$ dans le résultat précédent. Pour éviter ce problème, on fait une hypothèse supplémentaire sur $h_n$ de façon à rendre le biais négligeable. On obtient alors avec :

$$\hat{M}_2(x) := \frac{1}{n\hat{F}(h_n)} \sum_{i=1}^{n} K^2\left(\frac{d(X_i,x)}{h_n}\right), \hat{M}_1(x) := \frac{1}{n\hat{F}(h_n)} \sum_{i=1}^{n} K\left(\frac{d(X_i,x)}{h_n}\right)$$

le corollaire suivant :

**Corollaire 1** *Sous les hypothèses du Théorème 1, si* $h_n\sqrt{nF(h_n)} \to 0$, *alors :*

$$\frac{\hat{M}_1}{\sqrt{\hat{M}_2\hat{\sigma_\epsilon^2}}}\sqrt{n\hat{F}(h_n)}\left(\hat{r}(x) - r(x)\right) \to N(0,1),$$

*où* $\hat{\sigma_\epsilon^2}$ *est un estimateur convergent en probabilité vers* $\sigma_\epsilon^2$.

Ce corrolaire nous permet de donner des intervalles de confiance asymptotiques :

$$\left[\hat{r}(x) - t_{\frac{\alpha}{2}}\frac{\sqrt{\hat{M}_2\hat{\sigma_\epsilon^2}}}{\sqrt{n\hat{F}(h_n)}\hat{M}_1}, \hat{r}(x) + t_{\frac{\alpha}{2}}\frac{\sqrt{\hat{M}_2\hat{\sigma_\epsilon^2}}}{\sqrt{n\hat{F}(h_n)}\hat{M}_1}\right].$$

# Références

BOSQ D. (1996) *Nonparametric statistics for stochastic processes. Estimation and prediction.* Lecture Notes in Statistics, **110**, Springer-Verlag, New York.

BOSQ D. (2000) *Linear processes in function spaces.* Lecture Notes in Statistics, **149**, Springer-Verlag, New York.

FERRATY F., GOIA A. and VIEU P. (2002) Functionnal nonparametric model for time series : a fractal approach for dimension reduction. *Test*, **11**, (2) 317-344.

FERRATY F., MAS A. and VIEU P. (2006) Advances on nonparametric regression for fonctionnal data. *Pre-print*.

FERRATY F. and VIEU P. (2006) *Nonparametric modelling for functionnal data.* Springer-Verlag, New York.

MASRY E. (2005) Nonparametric regression estimation for dependent functional data : asymptotic normality. *Stochastic Process. Appl.*, **115**, (1), 155-177.

RIO E. (1995) About the Lindeberg method for strongly mixing sequences. *ESAIM*, **1**, 35-61.

RIO E. (2000) *Théorie asymptotique des processus aléatoires faiblement dépendants.* Mathématiques et applications, **31**, Springer-Verlag, Berlin

.

# Asymptotic normality of conditional quantile in the normed space under $\alpha$-mixing hypothesis

**M'hamed Ezzahrioui* and Elias Ould-Saïd**

* Adresse pour correspondance :
L.M.P.A. J. Liouville, Univ. du Littoral Côte d'Ôpale
BP 699, 62228 Calais, France
e-mail : m_ezzahri@yahoo.fr et ouldsaid@lmpa.univ-littora.fr

## Abstract

We consider the estimation of the conditional quantile function when the covariables take values in some abstract function space. The main goal of this paper is to establish the asymptotic normality and the almost completes convergence of the kernel estimator of the conditional quantile when the processes is assumed to be strongly mixing, some applications are given.

## Références

Bensaid, N. et 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.

Bensaid, N. and Oliveira, P.E. (2001). Histogram estimation of Radon-Nikodym derivatives for strong mixing data. *Statistics*, **35**, 569-592.

Berlinet, A., Gannoun, A. and Matzner-L$\phi$ber, E. (2001). Asymptotic normality of the convergent estimates of conditional quantiles. *Statistics*, **35**, 139-169.

Bertrand-Retali, M. (1977). Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Publ. Inst. Statist. Univ Paris*, **22**, 1-42.

Bollerslev, T. (1986). General autoregressive conditional heteroskedasticity. *J. Economt.* **31**, 307-327.

Bhattacharya, P.K., Gangopadhyay, A. (1990). Kernel and nearest neighbor estimation of conditional quantile, *Ann. Statist.* **18**, 1400-1415.

Cadre, B. (2001). Convergent estimators for the $L_1$-median of Banach valued random variables, *Statistics*, **35**, 509-521.

Chaudhuri, P. (1991a). Nonparametric estimates of regression quantiles and their local Bahadur representation, *Ann. Statist.* **19**, 760-777.

Chaudhuri, P. (1991b). Global nonparametric estimation of conditional quantile functions and their derivatives, *J. Multivariate Anal.* **39**, 246-269.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression, *Ann. Statist.* **25**, 715-744.

Dabo-Niang, S. (2004). Kernel density estimator in an infinite dimensional space with a rate of convergence, *Appl. Math. Lett.*, **17**, 381-386.

Engle R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987-1007.

Fan, J., Hu, T.C., and Truong, Y.K. (1994). Robust nonparametric function estimation, *Scand. J. Statist.* **21**, 433-446.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist. and Data Anal.*, 4, 545-564.

Ferraty, F. and Vieu, P. (2003). Curves discrimination : a nonparametric functional approach. *Comput. Statist. and Data Anal.*,**88**, 161-173.

Ferraty, F. and Vieu, P. 2004. Nonparametric models for functional data, with application in regression times series prediction and curves discrimination. *J. Nonparametric Statist.*, 16, 111-127.

Ferraty, F., Laksaci, A. and Vieu, P. (2005). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statist. Inf. for Stoch. Processes.* to appear.

Geffroy, J. (1974). Sur l'estimation d'une densité dans un espace métrique. C. R. Acad. Sér. A, **278**, 1449-1452.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for $\alpha$-mixing processes, *Ann. Inst. Statist. Math.* **52**, 459-470.

Jones, D.A. (1978). Nonlinear autoregressive processes. *Proc. Roy. Soc. London A,* **360**, 71-95.

Jones, M.C. and Hall, P. (1990). Mean squared error properties of kernel estimates of regression quantiles, *Statist. Probab. Letters* **10**, 283-289.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data : Asymptotic normality. Stoch. Proc. and their Appl. 115, 155-177.

Ozaki, T. (1979). Nonlinear time series models for nonlinear random vibrations. Technical report. Univ. of Manchester.

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

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis : Methods and Case Studies*, Springer, New-York.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles, *Statist. Probab. Lett.*, **7**, 407-412.

# Some ideas on depth for clustering with functional data.

**Antonio Cuevas, Manuel Febrero* & Ricardo Fraiman**

* Adresse pour correspondance :
Dpto de Estatística e I.O.
Facultad de Matemáticas
Avda Lope Gomez de Marzoa, s/n
15782 Santiago de Compostela - SPAIN
e-mail : mfebrero@usc.es

## Introduction

Clustering or unsupervised discrimination (or unsupervised learning in machine learning literature) is an important subject in statistics in many fields (Kaufman & Rousseeuw (1990)). Usually, its objectives are to find the locations and the number of clusters where the data are around. Really, this two problems are separate but it is tempting to try to solve both simultaneously. The choice of a clustering algorithm depends both on the type of data available and on the particular purpose. Broadly speaking there are two kinds of clustering algorithms, namely partitioning and hierarchical methods. A partitioning method constructs $k$ clusters where the $k$ is provided by the user, in three steps. First, $k$ centers are established, then every datum is classified in the group whose center is closer and finally the center are updated. This two last steps can be repeated until "convergence". Hierarchical methods deal with all values of $k$ in the same run, providing an evolutionary tree of clusters between $k = 1$ and $k = n$ using some rules to aglomerate (or separate) data from a group. An agglomeration algorithm begins with $k = n$ and a divisive algorithm begins with $k = 1$.
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)). This is the case of clustering algorithms. The usual tools in clustering are distances between objects and, in a functional setup, this is almost the only information someone can use.

But all the clustering algorithms are very vulnerable to outliers. The task, in multivariate data, of finding outliers, it is hard but affordable but in functional data it's clearly more complicated.
In the task of finding outliers the notion of depth can help. 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 for multivariate data. A depth measure provides an ordering of every datum from "center" to outward, so this ordering can be used to generalize usual concepts in univariate data like trimmed means.

In this work, a new procedure for clustering functional data is proposed along the following stages.

1. A depth measure is used in order to clean the original data. It will be important that this measure preserve the information about groups.
2. Using smooth bootstrap, an artificial sample is drawn from the cleaned data. The purpose of this stage is to fill the functional space avoiding the presence of "holes" in the data.
3. A clustering algorithm is then applied to the artificial sample, finding $k$ groups.
4. The original data is then classified in the group of the closest datum in the artificial sample.
5. The last three stages can be repeated to obtain a final estimator with bagging.

# Depth measures

As in classical univariate point estimation, we can look for alternative location estimators respect to sample mean, in order to get some idea about the "central value" of the population from which the sample of curves $x_1 = x_1(t), \ldots, x_n = x_n(t)$ has been drawn. In this work we will work with two estimators that are oriented to catch the bumps taking into account different aspects of the notion of "mode" with functional data.

As far as we know there is no widely accepted definition of mode for functional data. We suggest here a tentative notion which, in some sense, is oriented to select the trajectory most densely surrounded by other trajectories of the process. Given a kernel function $K : \mathbb{R} \to \mathbb{R}$ and a fixed tuning parameter $h$, we define

$$g(x; h) := g(x; h; x_1, \ldots, x_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right), \text{ and}$$
$$M(x_1, \ldots, x_n) = \max_i g(x_i; h).$$

The other choice is based on random projections (Cuesta-Albertos et al, 2005, Barrio et al, 2006 ). Given a process $X$ as well as an independent "direction process" $H$. Let $f[\langle h, X\rangle](.)$ the density function of the projection of $X$ along the direction given by the trajectory $h$ of $H$. Let $D[\langle h, X\rangle](t) = max\{f[\langle h, X\rangle](t)\}$. We define the *population deepest trajectory* as any function in the set

$$argmax_{\mu} E(D[\langle H, X\rangle](\langle H, \mu\rangle)).$$

We can define an empirical version of $D$ using

$$D_n[y_1, \ldots, y_n](t) = max\{f_n[y_1, \ldots, y_n](t)\}$$

where $f_n[y_1, \ldots, y_n](t)$ is a density estimation associated with the sample of real values $y_1, \ldots, y_n$ and defining the *sampling deepest function* to be the trajectory $x_{k_0}$ maximizing in $k$ the Functional Random Projection Depth

$$\frac{1}{N}\sum_{j=1}^{N} D_n[\langle h_j, x_1\rangle, \ldots, \langle h_j, x_n\rangle](\langle h_j, x_k\rangle)$$

where $h_1, \ldots, h_N$ is a random sample of trajectories of the process $H$.

Once a depth measure is chosen, we can rank the functional data according its depth measure, marking those with low rank to be considered as outliers. Also, for every depth measure, it is possible to construct additional functional $L$-estimators, as the $\alpha$-trimmed mean that can be defined as the average of the $100(1-\alpha)\%$ deepest functions in the sample.

# Clustering

For the construction of the clusters we use two different ideas.
The first one is basically the $k$-means algorithm. For this technique is very relevant to be sure that in the sample there are no outliers. If not, this algorithm usually identifies an outlier with the center of a cluster. An interesting alternative can be to substitute the mean by a robust location estimator.
The second is an adaptation of some ideas in Cuevas et al (2001). Let's define the empirical clusters as the connected components of $\{\hat{f}_n \geq c_n\}$ where

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right)$$

where $c_n$ is an estimator of

$$c_\alpha = max\{c : \mathbb{P}\{x : E\left(K\left(\frac{\|x - X\|}{h}\right)\right) \geq c\} \geq 1 - \alpha\}$$

Then, the data could be classified according to the empirical cluster they belong. Because for this algorithm it's important that there are no holes in the data, we propose "refill" the functional space drawing smoothed bootstrap observations $X_i^0 = X_i^* + Z_i$ with $X_i^0 \in \{\hat{f}_n \geq c_n\}$, $X_i^*$ obtained from ordinary bootstrap, $Z_i$ is a gaussian process with mean zero and covariance matrix $\Sigma_Z = r \times \Sigma$ being $r$ a small quantity. The bootstrap with functional data was previously succesfully applied in other contexts, see for example, Cuevas & Fraiman, 2004 and Cuevas et al, 2004. Then the algorithm is based on the idea of classifying the original observations according to the connected component they belong.

# Just an example

The following is just presented as an example of the complete procedure in a simple example just with two clusters. The data is generated by the model $X_n(t) = m_i(t) + Z_n(t)$ where $m_1(t) = 30x(1-x)^k$, $m_2(t) = 30x^k(1-x)$, $Z_i(t)$ is a gaussian process with mean zero and covariance fucntion $Cov(Z_i(t), Z_i(s)) = \exp(-|t-s|/0.3)$ and $k$ a parameter taking values from 1 to 1.5. With $k = 1$ the two groups are the same. With $k = 1.5$ the maximum distance between clusters is achieved. A sample of size 100 with equal number of elements in the groups is shown in Fig.1 and the corresponding cleaned data at 25% in Fig. 2 with $k = 1.5$. In the following table the percentage of correct classification is shown for 5 replicas of the original model for distinct $k$. Here, the mode was used as depth measure and the clustering was performed by $k$-means and the method of connected components.

| | Pct. Correct Classification | |
|---|---|---|
| $k$ | $k$-means | conn. comp. |
| 1 | 56 | 51 |
| 1.1 | 74 | 52 |
| 1.2 | 92 | 53 |
| 1.3 | 98 | 60 |
| 1.4 | 100 | 90 |
| 1.5 | 100 | 100 |

In this example, $k$-means method clearly outperfoms connected components. There are reasons for this behaviour. First, in this case the initial centers for $k$-means are selected quite optimal. Second, this example is not the example where connected components can do a better job than $k$-means.

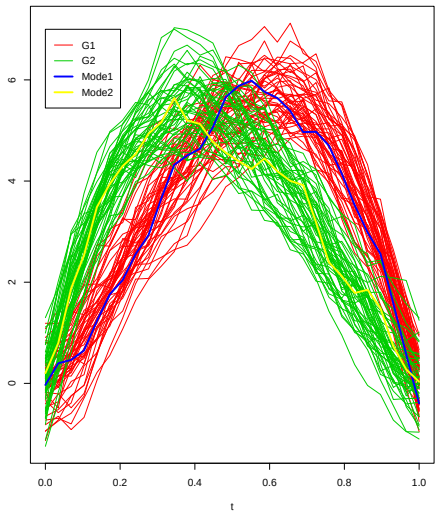


FIG. 3 – Sample data and modes

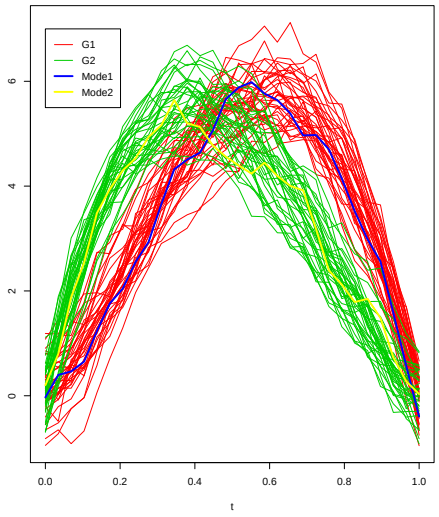


FIG. 4 – Cleaned data (25%) and modes

# References

Barrio, del E., Cuesta-Albertos, J.A., Fraiman, R. and Matran, C. (2006). The random projection method in goodness-of-fit tests for functional data. Preprint. Available at http ://personales.unican.es/cuestaj/.

Cuesta-Albertos, J.A., Fraiman, R. and Ransford, T. (2005). Random projection and goodness-of-fit tests in infinite dimensional spaces. Preprint. Available at http ://personales.unican.es/cuestaj/.

Cuevas, A., Febrero, M. and Fraiman, R. (2001). Cluster analysis : a further approach based on density estimation. Computational Statistics and Data Analysis, 36, 441 – 459.

Cuevas, A. and Fraiman, R. (2004). On the bootstrap methodology for functional data. In *Proceedings in Computational Statistics, COMPSTAT 2004*, J. Antoch ed., Physica-Verlag, pp. 127 – 135.

Cuevas, A., Febrero, M. and Fraiman, R. (2004). An anova test for functional data. Computational Statistics and data Analysis, 47, 111 – 122.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis.* Springer-Verlag.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data.* John Wiley & sons.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. J. R. Statist. Soc, B, 60, 4, 681 – 691.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional data analysis.* Springer-Verlag, New-York.

Ramsay, J.O. and Silverman, B.W. (2002). *Applied functional data analysis ; Methods and case studies.* Springer-Verlag, New-York.

# Large Nearest neighbor classification in infinite dimension

**Frédéric Cérou and Arnaud Guyader ***

* Adresse pour correspondance :
Université de Rennes 2
35 043 Rennes Cedex
e-mail : Arnaud.Guyader@uhb.fr

## Abstract

Let $X$ be a random element in a metric space $(\mathcal{F}, d)$, and let $Y$ be a random variable with value 0 or 1. $Y$ is calledthe class, or the label, of $X$. Assume $n$ i.i.d. copies $(X_i, Y_i)_{1 \leq i \leq n}$. The problem of classification is to predict the label of a new random element $X$. The $k$-nearest neighbor classifier consists in the simple following rule : look at the $k$ nearest neighbors of $X$ and choose 0 or 1 for its label according to the majority vote. If $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, Stone has proved in 1977 the universal consistency of this classifier : its probability of error converges to the Bayes error, whatever the distribution of $(X, Y)$. We show in this paper that this result is no more valid in general metric spaces. However, if $(\mathcal{F}, d)$ is separable and if a regularity condition is assumed, then the $k$-nearest neighbor classifier is weakly consistent.

## Références

Luc Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, 9(6) : 1310-1319, 1981.

Luc Devroye, Lázszló Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, Springer-Verlag, New-York, 1996.

Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4) : 595-645, 1977.

.

# Common functional principal components

**Aloïs Kneip ***

* Adresse pour correspondance :
Statistische Abteilung, Department of Economics
Universität Bonn, Adenauerallee 24
53113 Bonn, Germany
e-mail : akneip@uni-bonn.de

## Abstract

Functional principal component analysis (FPCA) based on the Karhunen-Loève decomposition has been successfully applied in many applications, mainly for one sample problems. In this paper we consider common functional principal components for two sample problems. Our research is motivated not only by the theoretical challenge of this data situation but also by the actual question of dynamics of implied volatility (IV) functions. For different maturities the log-returns of IVs are samples of (smooth) random functions and the methods proposed here study the similarities of their stochastic behavior. Firstly we present a new method for estimation of functional principal components from discrete noisy data. Next we present the two sample inference for FPCA and develop two sample theory. We propose bootstrap tests for testing the equality of eigenvalues, eigenfunctions, and mean functions of two functional samples, illustrate the test-properties by simulation study and apply the method to the IV analysis.

.

# Estimation of conditional distribution and conditional hazard function

**Frédéric Ferraty, Abbes Rabhi* and Philippe Vieu**

* Adresse pour correspondance :
Université Djillali Liabes, Sidi Bel Abbes
Algérie
e-mail : rabhi_abbes@yahoo.fr

## Résumé

Dans ce travail on se propose d'étudier la convergence uniforme presque complète d'un estimateur de la fonction du hasard conditionnelle dans le cas d'un processus fortement mélangeant sans présence de censure et avec censure aléatoire. Evidemment on étudie la distribution d'une variable aléatoire réelle conditionnée par une variable fonctionnelle, dont l'objectif sera l'estimation de la fonction de risque conditionnelle au moyen de la fonction de répartition conditionnelle et sa dérivée par la méthode du noyau. Cet étude exploite bien les propriétés de concentration sur de petites boules de la mesure de probabilité de la variable fonctionnelle explicative. Les différents résultats que nous allons présenter sont obtenus sous l'hypothèse d'indépendance et de mélange fort ou encore $\alpha$-mélange des données.

## A. Estimation avec données complètes

Dans de nombreuses études cliniques, le temps de survie (ou durée de survie) d'un patient est le critère principal d'évaluation thérapeutique, et pour des raisons diverses (décés d'un patient pour des causes extérieures à l'étude, abandon du patient en cours d'étude,...), il arrive que cette durée de survie ne puisse être observée complètement. On parle de censure. Nous admettrons que les sujets entrent de façon aléatoire et uniforme au cours du temps. Ansi, si la date d'analyse de l'essai est fixée à priori, le délai entre la date d'entrée et la date de fin d'expérience (date à la quelle on décide d'arrêter d'observer les sujets) est aléatoire. On qualifie alors la censure d'aléatoire. Dans la cadre de notre travail, nous nous limiterons au cas d'abscence de censure et le cas de la censure aléatoire

à droite, c'est à dire qu'en présence de censure, nous pourons seulement affirmer que "la durée de survie du patient est supérieure à une certaine valeur".

Le taux de hasard est défini comme étant la probabilité instantanée qu'une durée $X$ se termine. Plus précisément, $h$ est définie par :

$$h(x) = \lim_{\Delta x \longrightarrow 0} \frac{P\left(X \leq x + \Delta x / X \geq x\right)}{\Delta x}, \qquad (x > 0) \tag{32}$$

Il n'est pas difficile de voir que que le taux de hasard peut être réecrit à partir de la densité $f(.)$ et de la fonction de survie $S(.) = 1 - F(.)$ de $X$ :

$$h(x) = \frac{f(x)}{S(x)}. \tag{33}$$

Deux approches ont été proposées pour estimer le taux du hasard. La première remplace $f(x)$ et $S(x)$ dans l'expression (32) par leurs estimateurs $\hat{f}(x)$ et $\widehat{S}(x)$ respectivement, ce qui nous donne l'estimateur du taux de hasard par :

$$\widehat{h}(x) = \frac{\hat{f}(x)}{\widehat{S}(x)}. \tag{34}$$

La deuxième méthode est basée sur la relation entre le hasard cumulative et le taux de hasard où le hasard cumulative est définie par :

$$H(x) = \int_0^x h(u)du. \tag{35}$$

Nielson et Linton (1995) appelent ce type d'estimateurs par (**estimateur interne**). La relation entre le hasard cumulatif et le taux de hasard suggère que $\widehat{h}(x)$ peut être obtenue en lissant $\widehat{H}(x)$ en utilisant un noyau autrement dit :

$$\widehat{h}(x) = \int K_h(x - u)d\widehat{H}(u), \tag{36}$$

où $h$ est une largeure de fenêtre tel que $h \longrightarrow 0$ quand $n \longrightarrow \infty$.

Jusqu'à maintenant, l'intérêt à porter sur le taux de hasard va généralement dépendre de certaines covariables, par exemple, le temps de survie d'un patient va être affecté par plusieurs caractéristiques tel que l'âge et le genre. Le taux de hasard conditionnel de $X$ sachant $Z = z$ est définie par :

$$h^z(x) = \lim_{\Delta \longrightarrow 0} \frac{P\left(X \leq x + \Delta x / X > x, \quad Z = z\right)}{\Delta x} \qquad [x > 0]. \tag{37}$$

Ainsi le taux de hasard peut être écrit comme le taux de la densité conditionnelle $f^z(.)$ et la fonction de survie $S^z(.) = F^z(.)$ de $X$, c'est à dire :

$$h^z(x) = \frac{f^z(x)}{S^z(x)}. \tag{38}$$

Soit $(X, Z)$ un couple de variable aléatoire à valeur dans $\mathbb{R} \times \mathcal{F}$ où $\mathcal{F}$ est un espace semi-métrique muni dela distance $d(.;.)$. Ce travail est consacré au problème général de l'estimation de la fonction de hasard conditionnelle de la variable aléatoire réelle $X$ sachant la variable aléatoire fonctionnelle $Z$.

**1. Le cas indépendant**

Soit $(X_i, Z_i)_{1 \leq i \leq n}$ un échantillon d'observations du couple aléatoire $(X, Z)$. Ce travail est consacré au problème général de l'estimation de la fonction de hasard conditionnelle, dont on l'estimera au moyen de l'estimation non paramétrique de la la distribution conditionnelle ( *cond-cdf* ) de $X$ sachant $Z$ définie par : pour tout $z \in \mathcal{F}$, $\forall x \in \mathbb{R}$ $F^z(x) = P(X \leq x | Z = z)$ ainsi la fonction de hasard conditionnelle $X$ sachant $Z = z$ est définie par :

$$\forall z \in \mathcal{F},\ \forall x \in \mathbb{R} \quad h^z(x) = \frac{f^z(x)}{1 - F^z(x)} \tag{39}$$

où $F^z$ (resp. $f^z$) est la distribution conditionnelle (resp. la densité conditionnelle) de $X$ sachant $Z = z$ qu'on suppose qu'elle est absolument continuepar rapport à la mesure de Lebesgue sur $\mathbb{R}$ (resp. bornée). Dans toute la suite $z$ sera un point fixe dans $\mathcal{F}$, $N_z$ indiquera un voisinage fixe de $z$, $\mathcal{S}$ un sous ensemble compact dans $\mathbb{R}$.

**Notations générales et hypothèses**

Tous le long de notre étude, quand aucune confusion ne sera possible, on note par $A$ et $A'$ une certaine constante générique de $\mathbb{R}^{*+}$.

Par la suite, on fixe un point $z$ dans $\mathcal{F}$ dont on note $N_z$ un voisinage de ce point et on utilise la notation suivante :

$$B(z, h) = \{z' \in \mathcal{F} / d(z', z) < h\}$$

et on introduit les hypothèses suivantes :

(H1) $\forall z \in \mathcal{F},\ \forall h > 0,\ P\left(Z \in B(z, h)\right) = \phi_z(h) > 0,$

(H2) $\forall (x_1, x_2) \in \mathcal{S} \times \mathcal{S},\ \forall (z_1, z_2) \in N_z \times N_z,\ |F^{z_1\,(j)}(x_1) - F^{z_2\,(j)}(x_2)| \leq A_z \left(d(z_1, z_2)^{b_1} + |x_1 - x_2|^{b_2}\right),\ j = 0, 1,\ b_1, b_2 > 0,$

(H3) $\exists \nu < \infty,\ f^z(x) \leq \tau\ \forall (x, z) \in \mathbb{R} \times \mathcal{F},$

(H4) $\exists \beta > 0,\ F^z(x) \leq 1 - \beta,\ \ \forall (x, z) \in \mathbb{R} \times \mathcal{F}.$

**Convergence et vitesse de convergence presque complète**

L'objectif de cette partie est d'estimer la fonction de hasrd conditionnelle de $X$ sachant $Z = z$ et de donner des vitesses de convergences, naturellement on l'estimera au moyen de l'estimateur de la distribution conditionnelle *cond-cdf*

définie par :

$$\forall z \in \mathcal{F},\ \forall y \in \mathbb{R}\quad F^z(x) = P(X \leq x | Z = z) \tag{40}$$

et de sa dérivée définies par :

$$\widehat{F}^{z\,(j)}(x) = \frac{h_H^{-j} \sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right) H^{(j)}\left(h_H^{-1} x - X_i\right)\big)}{\sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right)}, \quad j = 0, 1 \tag{41}$$

où $K$ est un noyau, $H$ une fonction de répartition et $h_K = h_K, n$
(resp. $h_H = h_H, n$) est une suite de nombres réels positifs. Donc l'estimateur de la fonction conditionnelle du hasard est donnée par :

$$\widehat{h}^z(x) = \frac{\widehat{f}^z(x)}{1 - \widehat{F}^z(x)} \tag{42}$$

OULD-Said (1992) a présenté un estimateur relativement similaire mais dans le cas mélangeant avec censure dont $Z$ est une variable aléatoire réelle. Dans notre cas fonctionnelle où nos observations sont indépendantes et non censurées, on a besoin des conditions suivantes :

(H5) $\forall (x_1, x_2) \in \mathbb{R}^2$, $|H^{(j)}(x_1) - H^{(j)}(x_2)| \leq A|x_1 - x_2|$ et $\int |t|^{b_2} H^{(1)}(t)dt < \infty$,
$\exists \nu > 0,\ \forall j' \leq j + 1 \quad |x|^{1+\nu} |H^{(j)}(x)| \lim_{x \to \infty} |x|^{1+\nu} \left|H^{(j)}(x)\right| = 0, \quad j = 0, 1$

(H6) $K$ un noyau a support compact $(0, 1)$ vérifiant $0 < A_1 < K(t) < A_2 < \infty$,

(H7) $\lim_{n \to \infty} h_K = 0$ with $\lim_{n \to \infty} \dfrac{\log n}{n h_H^j \phi_x(h_K)} = 0, \quad j = 0, 1$

(H8) $\lim_{n \to \infty} h_H = 0$ with $\lim_{n \to \infty} n^a\, h_H = \infty \quad \forall a > 0.$

(H9) $H$ est le noyau d'une distribution cumultative dont le suport de $H^{(1)}$ est compact et $\forall l \geq j$, $H^{(l)}$ existe et bornée.

On établit le résultat suivant :

**Théorème 2** *Sous les hypothèses (H1)-(H9) on a :*

$$\sup_{x \in \mathcal{S}} |\widehat{h}^z(x) - h^z(x)| \longrightarrow 0, \qquad p.co. \tag{43}$$

**Théorème 3** *Sous les hypothèses (H1)-(H9) on a :*

$$\sup_{x \in \mathcal{S}} |\widehat{h}^z(x) - h^z(x)| = O\left(h_K^{b_1}\right) + O\left(h_H^{b_2}\right) + O\left(\sqrt{\frac{\log n}{n\, h_H\, \phi_z(h_K)}}\right), \qquad p.co. \tag{44}$$

**2. Le cas dépendant**

Pour pouvoir étendre au cas dépendant les résultats obtenus dans le cas indépendant. Nous allons adopter certaines hypothèses sur le processus $(X_i,\ X_j)_{i\in\mathbb{N}}$. Soient :

$$Z_i :\ (\Omega,\ \mathcal{A},\ P) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) X_i :\ (\Omega,\ \mathcal{A},\ P) \longrightarrow (\mathcal{F}, \Im)$$

où $\mathcal{F}$ est muni d'une semi-métrique $d,\ i \in \mathbb{N}$. $(X_i, Z_i)_{i\in\mathbb{N}}$, on se propose d'estimer la fonction de hasard conditionnelle de $X$ sachant $Z = z$.

Par la suite nous notons $(X_i, Z_i) = Y_i$ et l'on considère que $(Y_i)_i$ est algébriquement $\alpha$-mélangeant.

Ainsi, on modélise la dépendance par la notion de mélange fort appelée encore $\alpha$-mélange. Par la suite on introduit les hypothèses suivantes.

**Notations générales et hypothèses**

(H0) La suite des variables aléatoires $(X_i, Z_i)$ est $\alpha$-mélangeante dont le coefficient de mélange vérifie :

$$\exists a,\ c \in \mathbb{R}_+^* :\ \forall n \in \mathbb{N}\ \ \alpha(n) \leq cn^{-a} \tag{45}$$

(H'2) $0 < \sup_{i\neq j} P\left((Z_i, Z_j) \in B(z,h) \times B(z,h)\right) = O\left(\frac{(\phi_z(h))^{(a+1)/a}}{n^{1/a}}\right)$.

**Convergence et vitesse de convergence presque complète**

L'objectif de cette partie est d'estimer la fonction de hasard conditionnelle de $X$ sachant $Z = z$ et de donner des vitesses de convergences, naturellement on l'estimera comme dans le cas indépendant au moyen de l'estimateur de la distribution conditionnelle *cond-cdf* définie en (40) et de sa dérivée définies en (41), dont l'estimateur de la fonction de hasard conditionnelle est donnée en (42).

**Théorème 4** *Sous les hypothèses (H'2), (H0)-(H9) et si*

$$\exists \eta > 0,\ Cn^{\frac{3-a}{a+1}+\eta} \leq \phi_z(h_K) \leq C' n^{\frac{1}{1-a}} \tag{46}$$

*où a vérifie l'inégalité suivante* $a > (5+\sqrt{17})/2$ *alors on a :*

$$\sup_{x\in\mathcal{S}} |\widehat{h}^z(x) - h^z(x)| \longrightarrow 0, \qquad p.co. \tag{47}$$

**Théorème 5** *Sous les hypothèses (H'2), (H0)-(H9) et si les deux inégalités suivantes sont satisfaites*

$$\exists \eta > 0,\ Cn^{\frac{3-a}{a+1}+\eta} \leq h_H \phi_z(h_K) \quad et \quad \phi_z(h_K) \leq C' n^{\frac{1}{1-a}} \tag{48}$$

*on a :*

$$\sup_{x\in\mathcal{S}} |\widehat{h}^z(x) - h^z(x)| = O\left(h_K^{b_1}\right) + O\left(h_H^{b_2}\right) + O\left(\sqrt{\frac{\log n}{n\, h_H\, \phi_z(h_K)}}\right), \qquad p.co. \tag{49}$$

# B. Estimation avec données censurées

Dans cette section, nous adopterons les mêmes notations que celles utilisées en analyse de survie.
Dans des domaines aussi variés que la médecine, l'épidémologie ou l'industrie, il arrive souvent qu'on ne dispose pas de toutes les données. Nous disons alors que nous sommes en présence de *censures*. Les méthode standards d'estimation ne sont plus applicables. Le schéma classique de censure se présente comme suit :
Soit $X$ une variable aléatoire positive représentant le temps (la durée) de survie ou de défaillance d'un individu ou d'un sujet prenant part à une étude clinique ou participant à une expérience médicale de fonction de répartition $F$ et de densité $f$, $Z$ la variable aléatoire fonctionnelle explicative à valeurs dans $(\mathcal{F}, d(.;.))$ de fonction de répartition $G$ et de densité $g$ et $C$ la variable aléatoire positive dite de censure (la variable aléatoire $X$ est censurée à droite par $C$) de fonction de répartition $F_1$ et de densité $f_1$. Ainsi les variables aléatoires observées sont $T_i = \min(X_i,\ C_i)$, $\Delta_i = I_{X_i \leq C_i}$ et $Z_i$, où $1 \leq i \leq n$. La fonction de survie
$S(x) = P(X > x)$ a bénéficie d'une litérature abondante. La fonction de survie $S(x)$ n'est autre que le complémentaire de la fonction de répartition. Comment interprêter la fonction du risque [1] ? En fait c'est la dérivée d'une probabilité que la durée soit comprise entre $x$ et $dx$, sachant que l'on atteint le période $x$. Plus pratiquement il s'agit d'un taux instantané de sortie de l'état à la date $x$. La courbe de survie prend une signification particulière donnée par :

$$S(x) = \exp\left[-\int_0^x h(u)du\right]$$

Nous allons maintenant présenter et détailler nos différentes contributions à ce thème. Ces contributions concernent l'estimation de la fonction de hasard conditionnelle et sa vitesse de convergence.

**Estimation de la fonction de hasard**
L'objectif de cette parie est d'estimer la fonction de hasrd conditionnelle de $X$ sachant $Z = z$ pour des données indépendantes et mélangeantes avec censure aléatoire, naturellemnt on l'estimera au moyen de l'estimateur de le distribution conditionnelle *cond-cdf* et de sa dérivée définies par 41 ainsi l'estimateur de la fonction conditionnelle du hasard est donné par (42). Pour pouvoir étendre au cas des données avec censure les résultats obtenus dans le cas des données complètes, nous allons adopter certaines hypothèses sur le processus $(X_i,\ Z_i)_{i\in\mathbb{N}}$. Soient :

$$X_i :\ (\Omega,\ \mathcal{A},\ P) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) Z_i :\ (\Omega,\ \mathcal{A},\ P) \longrightarrow (\mathcal{F}, \Im)$$

où $\mathcal{F}$ est muni d'une semi-métrique $d,\ i \in \mathbb{N}$. On note $\varphi = f(1 - F_1)$ où $\varphi^z$ n'est autre que la densité commune conditionnelle des variables alétoires $(T_i,\ \Delta_i = 1)$

[1] $h(x)$ est le risque instantané de décés, il est traduit parfois à tors par fonction de hasard car en englais hazard veut dire risque.

qu'on suppose qu'elle est bornée ($\varphi^z(t) < \alpha,\ \forall t \in \mathbb{R}_+$) où $\alpha \in \mathbb{R}$ et $L^z$ la fonction de répartition conditionnelle de $T = \min(X,\ C)$ sachant $Z = z$ qu'on suppose qu'elle est absolument continue par rapport à la mesure de Lebesgue. Ainsi on définit la fonction de hasard conditionnelle par :

$$\forall z \in \mathcal{F},\ \forall t \in \mathbb{R}_+ \quad h^z(t) = \frac{f^z(t)}{1 - F^z(t)} = \frac{\varphi^z(t)}{1 - L^z(t)}, \quad L^z(t) < 1 \tag{50}$$

avec $L^z(t) = (1 - F^z(t))\,(1 - F_1^z(t))\,,\ \varphi^z(t) = f^z(t)\,(1 - F_1^z(t))$. Ainsi on estime $h^z$ au moyen des estimateurs de $\varphi^z$ et $L^z$ où l'estimateur de $L^z$ est donné par :

$$\text{Ł}^z(t) = \frac{\displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right) H\left(h_H^{-1}(t - T_i)\right)}{\displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right)} \tag{51}$$

et $\varphi^z$ est estimé par

$$\widehat{\varphi}^z(t) = \frac{h_H^{-1} \displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right) \Delta_i H\left(h_H^{-1}(t - T_i)\right)}{\displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d(z, Z_i)\right)} \tag{52}$$

Ainsi l'estimteur de la fonction de hasard est donné par :

$$\widehat{h}^z(t) = \frac{\widehat{\varphi}^z(t)}{1 - \text{Ł}^z(t)}. \tag{53}$$

**Convergence et vitesse de convergence presque complète : cas indépendant**
Nous supposons qu'on dispose de $n$ couples d'observations indépendantes $(T_i,\ \Delta_i)$ où $T_i = \min(X_i,\ C_i)$ et $\Delta_i = I(T_i \leq C_i)$. Nous supposons de plus que les variables $X$ et $C$ sont indépendantes (condition d'identifiabilité). Si on outre une covariable $Z$ est associée à $X$, $Z$ représente par exemple l'âge, le sexe, la pression artérielle ou le taux de cholestérol, le modèle de censure se présente sous forme de triplets $(Z_1,\ T_1,\ \Delta_1), \ldots, (Z_n,\ T_n,\ \Delta_n)$ indépendants et identiquement distribués. Pour l'identifiabilité, nous supposons que $X$ est indépendante de $C$ sachant $Z$, ainsi les variables aléatoires observées sont $T_i$, $\Delta_i$ et $Z_i$. En reprenant les mêmes arguments développer dans le cas des donnée complète dont on introduit les conditions ci-dessous pour pouvoir étendre les résultats obtenus dans le cas d'absence de censure.

(H10) $\forall (t_1, t_2) \in \mathcal{S}\times\mathcal{S},\ \forall (z_1, z_2) \in N_z \times N_z, |L^{z_1}(t_1) - L^{z_2}(t_2)| \leq A_z \left(d(z_1, z_2)^{b_1} + |t_1 - t_2|^{b_2}\right)$,

(H11) $\exists \mu < \infty,\ \varphi(t,z) < \mu,\ \forall (t,z) \in \mathbb{R}_+ \times \mathcal{F}$,

(H12) $\exists \eta > 0,\ L^z(t) \leq 1 - \eta,\ \ \forall (t,z) \in \mathbb{R}_+ \times \mathcal{F}$.

On établit le résultat suivant :

**Théorème 6** *Sous les hypothèses (H1)-(H12) on a :*

$$\sup_{t \in \mathcal{S}} \left| \widehat{h}^z(t) - h^z(t) \right| \longrightarrow 0, \qquad p.co. \tag{54}$$

*et*

$$\sup_{t \in \mathcal{S}} |\widehat{h}^z(t) - h^z(t)| = O\left(h_K^{b_1}\right) + O\left(h_H^{b_2}\right) + O\left(\sqrt{\frac{\log n}{n\, h_H\, \phi_z(h_K)}}\right), \qquad p.co. \tag{55}$$

**Convergence et vitesse de convergence presque complète : cas dépendant**
On rappelle que l'on observe ici l'échantillon $(T_i, \quad \Delta_i)_{1 \leq i \leq n}$ où $T_i = \min(X_i,\ C_i)$ avec $C$ censure aléatoire à droite et $\Delta_i = I_{X_i \leq C_i}$. On s'intéresse cependant aux estimateurs de la fonction de survie conditionnelle de la variable aléatoire $T$ sachant $Z = z$, dont on introduit l'hypothèse suivant :

(H'0) Le processus $(T_i,\ \Delta_i,\ \ Z_i)_{i \in \mathbb{N}}$est strictement stationnaire et $\alpha$-mélangeant dont le coefficient de mélange vérifie l'équation (45).

**Théorème 7** *Sous les hypothèses (H'0), (H'2) (H1)-(H12) on a :*

$$\sup_{t \in \mathcal{S}} \left| \widehat{h}^z(t) - h^z(t) \right| \longrightarrow 0, \qquad p.co. \tag{56}$$

*et*

$$\sup_{t \in \mathcal{S}} |\widehat{h}^z(t) - h^z(t)| = O\left(h_K^{b_1}\right) + O\left(h_H^{b_2}\right) + O\left(\sqrt{\frac{\log n}{n\, h_H\, \phi_z(h_K)}}\right), \qquad p.co. \tag{57}$$

# Références


Fan, J. and I. Gijbels (1996). Local Polynomial Modelling and Its Applications. London, Chapman and Hall.

Fan, j., Q. Yao, and H. Tong (1996). Estimation of Conditional densities and Sensitivity Measurs in Nonlinear Dynamical Systems. *Biometrika*, **83**, 189-206.

Fan, J. and Q. Yao (2003). Nonlinear Time Series. Nonparametric and Parametric Methods. *Springer Verlag.*

Ferraty, F. Laksaci, A and Vieu, P. (2004) *Estimating some characteristics of the conditional distribution in nonparametric functional models.*

Gefeller, O. and P. Michels (1992). A review on Smoothing Methods for the Estimation of the Hazard Rate based on Kernel Functions, in Dodge, Y. and J. Whittaker (eds) *Computational Statistics*, Physica-Verlag, Switzerland, 459-464.

Gonzàlez-Manteiga, W. , R. Cao, and J. S. Marron (1996). Bootstrap Selection of the Smoothing Parameter in Nonparametric Hazard Rate Estimation. *Journal of American Statistical Association*, **91**, 1130-1140.

Grambbsch, P.M. and T.M. Therneau (1994). Proportional Hazard Tests and Diagnostics Based on Weighted Residuals. *Biometrica*, **81**, 515-526.

Gray, R.J, (1996). Hazard Rate Regression Using Ordinary Nonparametric Regression Smoothers. *Journal of Computational and Graphical Statistics*, **5**, 190-207.

Hansen, B.E (2004). Nonparametric Estimation of Smooth Conditional Distribution Functions. *Workig paper University of Wisconsin-Madison.*

Hyndman, R.J., D.M. Bashtannyk, and G.K. Grunwald (1996). Estimating and Visualizing Conditional Density Estimators. *Journal of Computational and Graphical Statistics*, **5**, 315-336.

Klein, J.P. and M.L. Moeschberger (1997). Survival Analysis : Techniques for Censored and Truncated Data. *Springer Verlag.*

Kooperberg, C., C.J. Stone and Y.K. Truong (1995). Hazard Regression. *Journal of the American Statistical Association*, **90**, 78-94.

Lecoutre, J-P. andOuld-Said, E. (1992). Estimation of conditional density and conditional hazard function from strong-mixing and censored processes. *Comptes Rendus Acad. Sci. Paris, Ser. I*, **331**, 295-300.

Li, G. and H. Doss (1995). An Approach to Nonparametric Regression for Life History Data Using Local Linear Fitting. *Annal of Statistics*, **23**, 787-823.

Liero, H. (2004). Teting the Hazard Rate Part I, Institute of Mathematics and University of Potsdam, mimeo.

Linton, O.B., J.P. Nielsen, and S. van de Geer (2003). Estimating Multiplicative and Additive Hazard Functions by Kernel Methods. *Annals of Statistics*, **31**, 464-492.

Lo, S.-H., Y.P. Mack, and J.L. Wang (1989). Density and Hazard Rate Estimation for Censored Data via Strong Representation of the Kaplan-Meier Estimator. *Probability theory Related Fields*, **80**, 461-472.

Müller, H.G and J.L. Wang (1990). Locally Adaptive Hazard Smoothing. *Probability theory and Related Fields*, **85**, 523-538.

Müller, H.G and J.L. Wang (1994). Hazard Rate Estimation Under Random Censoring with Varying Kernels and Bandwiths. *Biometrics*, **50**, 61-76.

Nielson, J.P., O.B. Linton (1995). Kernel Estimation in a Nonparametric Marker Dependent Hazard Model. *Annals of Statistics*, **23**, 1735-1748.

Nielson, J.P., O.B. Linton, and P.J. Bickel (1998). On a Semiparametric Survival Model with Flexated Covariate Effect. *Annals of Statistics*, **26**, 215-241.

Orbe, J., E. Ferreira, and V. Nùñez-Antòn (2002). Comparing Proportional Hazards Accelerated Failure Time Models for Survival Analysis. *Statistics in Medicine*, **21** , 3492-3510.

Padgett, W.J. (1988). Nonparametric Estimation of Density and Hazard Rate Functions when Samples are Censored, P.R. Krishnaiah and C.R. Rao (eds). *Handbook of Statistics*, **7**, 313-331. Elsevier Science Publishers.

Patil, P.N. (1993). Bandwith Choice for Nonparametric Hazard Rate Estimation. *Journal of Statistical Planning and Inference*, **35**, 15-30.

Patil, P.N. (1993b). On the Least Square Cross-Validation Bandwith in Hazard Rate Estimation. *Anal of Statistics*, **21**, 1792-1810.

Rice, J. and M. Rosenblatt (1976). Estimation of the Log Survivor Function and Hazard Function. *Sankhya Series A*, **38**, 60-78.

Sarda, P. and P. Vieu (19991). Smoothing Parameter Selection in Hazard Estimation. *Statistics and Probability Letters*, **11**, 429-434.

Staph, (2002). Proceedings of the working group on Functional Statistics : 2001-2002. *Techn. report, Labo. Statist. Proba. Toulouse, France*, **2002-12**. (Available on line at www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html).

Tanner, M.A. and W.H. Wong (1983). The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Methods. *Annals of Statistics*, **11**, 989-993

Tanner, M.A. and W.H. Wong (1984). Data-Based Nonparametric Estimation of the Hazard Function with Applications to Model Diagnostics and Exploratory Analysis. *Journal of the American Statistical Association*, **79**, 174-182.

Therneau, T.M. and P.M. Grambsch (2000). Modeling Survival Data : Extending the Cox Model, Springer Verlag.

Van Keilegom, I. and N. Veraverbeke (2001). Hazard Rate Estimation in Nonparametric Regression with Censored Data. *Ann. Inst. Statist. Math*, **53**, 730-745.

# Une méthode SIR multivariée ($PMS_\alpha$) en présence de covariables qualitatives

**Benoît Liquet et Jérôme Saracco***

*Adresse pour correspondance :
Institut de Mathématiques de Bourgogne, UMR CNRS 5584,
Université de Bourgogne,
9 avenue Alain Savary, 21 078 Dijon Cedex, France
e-mail : Jerome.Saracco@u-bourgogne.fr et benoit.liquet@upmf-grenoble.fr

## Résumé

Nous considérons ici un modèle semi-paramétrique de régression prenant en compte non seulement une covariable quantitative $p$-dimensionnelle, notée $X$, mais aussi un prédicteur qualitatif $Z$. Ce modèle comprend une réduction de la dimension de $X$ par le biais de $K$ indices $X'\beta_k$. La variable à expliquer $Y$ peut être ou bien réelle ou bien $q$-dimensionelle. Nous proposons une approche fondée sur la méthode $SIR_\alpha$ et la méthode du "Pooled Marginal Slicing" (combinaison de tranchages marginaux) afin d'estimer la partie paramétrique du modèle, à savoir le sous-espace linéaire engendré par les vecteurs $\beta_k$. Cette approche, contrairement aux travaux de Chiaromonte et al. (2002), ne nécessite pas d'hypothèse d'homoscédasticité pour les variances conditionnelles de $X$ sachant $Z$. Nous avons établi la convergence en probabilité à la vitesse $\sqrt{n}$ de notre estimateur. Nous illustrons enfin, sur des simulations, les qualités numériques de cet estimateur.

## Abstract

We consider a semiparametric regression model involving both $p$-dimensional quantitative covariable $X$ and categorical predictor $Z$, and including a dimension reduction of $X$ via $K$ indices $X'\beta_k$. The dependent variable $Y$ can be real or $q$-dimensional. We propose an approach based on $SIR_\alpha$ and Pooled Marginal Slicing methods in order to estimate the space spaned by the $\beta_k$'s. We establish $\sqrt{n}$-consistency of the proposed estimator. Simulation studies show the numerical qualities of our estimator.

# Introduction

Dans cette communication, nous considérons le modèle semi-paramétrique suivant dans lequel la variable à expliquer $Y$ est à valeurs dans $\Re^q$ ($q \geq 1$), la covariable quantitative $X$ est $p$-dimensionnelle ($p > 1$) et le prédicteur qualitatif $Z$ prend $L$ niveaux : pour $l = 1, \ldots, L$,

$$Y = \begin{cases} Y_1^{(l)} = g_1^{(l)}(\beta_1' X, \ldots, \beta_K' X, \varepsilon_1^{(l)}) & \text{si } Z = l, \\ \vdots & \\ Y_q^{(l)} = g_q^{(l)}(\beta_1' X, \ldots, \beta_K' X, \varepsilon_q^{(l)}) & \text{si } Z = l. \end{cases} \tag{58}$$

Une réduction de dimension de la covariable $X$ est obtenue au moyen des $K$ indices $\beta_k' X$. Le prédicteur qualitatif $Z$ est une covariable catégorielle qui n'est pas inclue dans la réduction de dimension. Cette covariable $Z$ peut représenter une ou plusieurs variables qualitatives qui permettent d'identifier $L$ sous-populations.

Pour simplifier, dans la suite, nous supposerons $K = 1$. Nous nous focaliserons alors sur la caractérisation et sur l'estimation de la direction EDR ("effective dimension reduction") $b$ colinéaire à $\beta := \beta_1$. Comme dans le cadre usuel des méthodes de type SIR ("sliced inverse regression", que l'on peut traduire par "régression inverse par tranches") introduite par Li (1991), une condition de linéarité est ici nécessaire pour chaque sous-population :

$$\forall v \in \Re^p, \quad \mathbf{E}[v'X|\beta' X, Z = l] \text{ est linéaire en } \beta' X \text{ pour chaque } l = 1, \ldots, L. \tag{59}$$

Nous définissons un estimateur de l'espace EDR reposant une approche de type $\text{PMS}_\alpha$ (méthode du "pooled marginal slicing", que l'on peut traduire par "combinaison de tranchages marginaux", fondée sur l'approche $\text{SIR}_\alpha$, combinaison des approches SIR-I et SIR-II, voir Barreda et al., 2006, pour plus de détails). Nous proposons de combiner les matrices "marginales" d'intérêt obtenues pour chaque composantes $Y_j$ de $Y$ et pour chaque niveau $l$ de $Z$. Ainsi, la version sur population de la matrice "finale" d'intérêt est la suivante :

$$\mathcal{M}_{q,L}^P = \sum_{j=1}^{q} w_q^{(j)} \left\{ \sum_{l=1}^{L} w_L^{(l)} \left(\Sigma^{(l)}\right)^{-1} M_{\alpha^{(j,l)}}^{(j,l)} \right\}, \tag{60}$$

où les matrices $M_{\alpha^{(j,l)}}^{(j,l)}$ sont les matrices $M_\alpha$ correspondant à la sous-population $l$ pour la composante $Y_j$ de $Y$, c'est-à-dire les matrices utilisées par la méthode $\text{SIR}_\alpha$ et définies par les couples $(X, Y_j)$ sachant $Z = l$. Les poids $\{w_L^{(l)},\ l = 1, \ldots, L\}$ sont les probabilités des événements $Z = l$, et les poids $\{w_q^{(j)},\ j = 1, \ldots, q\}$ sont des poids positifs tels que $\sum_{j=1}^q w_q^{(j)} = 1$. Le paramètres $\alpha$ de chaque matrice $M_\alpha$ peut être choisi indivuellement et est alors noté $\alpha^{(j,l)}$. Dans la suite, nous utiliserons, pour simplifier, l'équi-pondération $\{w_j^{(q)} = \frac{1}{q},\ j = 1, \ldots, q\}$.

Comme dans Chiaromonte et al. (2002), nous allons d'abord considérer l'hypothèse simplificatrice d'homoscédacticité suivante :

$$\Sigma^{(l)} = \Sigma^\star, \quad l = 1, \ldots, L, \tag{61}$$

C'est à dire que l'on suppose que toute les structure de covariance de $X$ sachant $Z = l$ sont les mêmes pour les $L$ sous-populations. Nous traiterons ensuite du cas hétéroscédactique. Dans chacun des cas, nous décrirons la version sur population et sur échantillon de la matrice d'intérêt (60).

# Cas homoscédastique

Lorsque $\Sigma^{(l)} = \Sigma^\star$ pour chaque sous-population $l = 1, \ldots, L$, la matrice $\mathcal{M}_{q,L}^P$ définie en (60) peut alors s'écrire :

$$(\Sigma^\star)^{-1} M_{q,L}^P, \tag{62}$$

où $M_{q,L}^P = \sum_{j=1}^q w_q^{(j)} \left\{ \sum_{l=1}^L w_L^{(l)} M_{\alpha^{(j,l)}}^{(j,l)} \right\}$. Clairement, sous la condition (59), le vecteur propre associé à la plus grande valeur propre de $(\Sigma^\star)^{-1} M_{q,L}^P$ est une direction EDR.

**Version sur échantillon.** Soit un échantillon i.i.d. $\mathcal{S} = \{(X_i, Z_i, Y_i),\ i = 1, \ldots, n\}$. Afin d'obtenir un estimateur de la matrice d'intérêt (62), l'idée usuelle des approches de type SIR est de remplacer les versions théoriques de tous les moments conditionnels de $X$ sachant $Y$ par leurs versions empiriques.

Soient $\widehat{\Sigma}^{(l)}$ les matrices de variances de $X$ dans chaque sous-échantillon

$$\mathcal{S}^{(l)} = \{(X_i, Y_i) \text{ tel que } Z_i = l\}$$

correspondant à la sous-population $l$. Un estimateur de la matrice "commune" de covariance est donné par : $\widehat{\Sigma}^\star = \sum_{l=1}^L \frac{n_l}{n} \widehat{\Sigma}^{(l)}$, où $n_l$ est la taille du sous-échantillon $\mathcal{S}^{(l)}$.

Introduisons le sous-échantillon $\mathcal{S}^{(j,l)} = \{(X_i, Y_{j,i}) \text{ tel que } Z_i = l\}$. Nous supposons que, pour chaque sous-population $l$, le support de chaque composante $Y_j$ de $Y$ est partionné en $H^{(j,l)}$ tranches $s_1^{(j,l)}, \ldots, s_h^{(j,l)}, \ldots, s_{H^{(j,l)}}^{(j,l)}$. Soit $n_h^{(j,l)}$ le nombre d'observations dans la tranche $h$ du sous-échantillon $\mathcal{S}^{(j,l)}$ pour la composante $Y_j$. Pour chaque sous-échantillon $\mathcal{S}^{(j,l)}$, nous calculons les moyennes et les matrices de variance par tranches : pour $h = 1, \ldots, H^{(j,l)}$, $j = 1, \ldots, q$ et $l = 1, \ldots, L$,

$$\bar{x}_h^{(j,l)} = \frac{1}{n_h^{(j,l)}} \sum_{i \in \mathcal{S}^{(j,l)}} X_i \mathbf{I}\left[ Y_{j,i} \in s_h^{(j,l)} \right],$$

$$\text{et } \widehat{V}_h^{(j,l)} = \frac{1}{n_h^{(j,l)}} \sum_{i \in \mathcal{S}^{(j,l)}} (X_i - \bar{x}_h^{(j,l)})(X_i - \bar{x}_h^{(j,l)})' \mathbf{I}\left[Y_{j,i} \in s_h^{(j,l)}\right],$$

où $\mathbf{I}[.]$ est la fonction indicatrice. Définissons ensuite $\widehat{\overline{V}}^{(j,l)} = \sum_{h=1}^{H^{(j,l)}} \frac{n_h^{(j,l)}}{n_l} \widehat{V}_h^{(j,l)}$. Les matrices $M_I^{(j,l)}$ et $M_{II}^{(j,l)}$ sont alors estimées de la manière suivante :

$$\widehat{M}_I^{(j,l)} = \sum_{h=1}^{H^{(j,l)}} \frac{n_h^{(j,l)}}{n_l} \left(\bar{x}_h^{(j,l)} - \bar{x}^{(j,l)}\right) \left(\bar{x}_h^{(j,l)} - \bar{x}^{(j,l)}\right)'$$

et

$$\widehat{M}_{II}^{(j,l)} = \sum_{h=1}^{H^{(j,l)}} \frac{n_h^{(j,l)}}{n_l} \left(\widehat{V}_h^{(j,l)} - \widehat{\overline{V}}^{(j,l)}\right) (\widehat{\Sigma}^\star)^{-1} \left(\widehat{V}_h^{(j,l)} - \widehat{\overline{V}}^{(j,l)}\right)'.$$

Finalement, nous définissons les estimateurs de $M_{\alpha^{(j,l)}}^{(j,l)}$ et $M_{q,L}^P$ ainsi :

$$\widehat{M}_{\alpha^{(j,l)}}^{(j,l)} = (1 - \alpha^{(j,l)}) \widehat{M}_I^{(j,l)} (\widehat{\Sigma}^\star)^{-1} \widehat{M}_I^{(j,l)} + \alpha^{(j,l)} \widehat{M}_{II}^{(j,l)}$$

et

$$\widehat{M}_{q,L}^P = \frac{1}{q} \sum_{j=1}^{q} \left\{ \sum_{l=1}^{L} \frac{n_l}{n} \widehat{M}_{\alpha^{(j,l)}}^{(j,l)} \right\}.$$

La direction EDR estimée $\hat{b}$ est alors le vecteur propre associé à la plus grande valeur propre de $(\widehat{\Sigma}^\star)^{-1} \widehat{M}_{q,L}^P$.

**Résultat symptotique.** Lorsque $n_h^{(j,l)} \to +\infty$ pour $n \to +\infty$, nous montrons la convergence en probabilité à la vitesse $n^{1/2}$ dela direction EDR estimée $\hat{b}$ vers la vraie direction EDR $b$. La normalité asymptotique de $\sqrt{n}\left((\widehat{\Sigma}^\star)^{-1} \widehat{M}_{q,L}^P - (\Sigma^\star)^{-1} M_{q,L}^P\right)$ et des ces éléments propres (projecteurs propres et vecteurs propres associés aux valeurs propres non nulles de $-(\Sigma^\star)^{-1} M_{q,L}^P$) peut aussi être obtenue comme dans Gannoun et Saracco (2003) ou Saracco (2005).

# Cas hétéroscédastique

Nous considérons ici la décomposition aux valeurs propres de la matrice $\mathcal{M}_{q,L}^P$ définie en (60). Il est important de noter ici qu'il n'y a aucune raison pour qu'il existe un produit scalaire spécifique tel que cette matrice soit symétrique ou définie positive. Il est donc possible de trouver des éléments propres complexes. Cependant, le résultat crucial est que, pour chaque matrice $\left(\Sigma^{(l)}\right)^{-1} M_{\alpha^{(j,l)}}^{(j,l)}$ (avec $j = 1, \ldots, q$ et $l = 1, \ldots, L$), le vecteur propre associé à la plus grande valeur propre est une direction EDR. Ainsi, vu que la matrice $\mathcal{M}_{q,L}^P$ est une combinaison

convexe des matrices $\left(\Sigma^{(l)}\right)^{-1} M^{(j,l)}_{\alpha^{(j,l)}}$, il est direct de voir qu'il existe un vecteur propre $b$ qui est une direction EDR associée à une valeur propre positive réelle $\lambda$. D'un point de vue algébrique, il n'y a aucune garantie que la valeur propre correspondante soit la plus grande (en module). On peut trouver des cas pathologique, un exemple est décrit dans Liquet et Saracco (2006). Malheureusement, nous ne connaissons pas de caractérisation de tels cas pathologiques, ni de conditions nécessaires permettant d'éviter de tels cas. C'est un problèmes ouvert. Cependant, on a pu remarquer sur nos simulations que cette pathologie n'a jamais été rencontrée.

**Version sur échantillon.** Nous estimons la matrice $\mathcal{M}^P_{q,L}$ en substituant les moments empiriques aux moments théoriques correspondants :

$$\widehat{\mathcal{M}}^P_{q,L} = \frac{1}{q}\sum_{j=1}^{q}\left\{\sum_{l=1}^{L}\frac{n_l}{n}\left(\widehat{\Sigma}^{(l)}\right)^{-1}\widehat{M}^{(j,l)}_{\alpha^{(j,l)}}\right\}, \tag{63}$$

où la modification majeure est que la matrice de covariance "commune" estimée $\widehat{\Sigma}^\star$ est maintenant remplacée par les matrices de covariances marginales $\widehat{\Sigma}^{(l)}$ des sous-échantillons $\mathcal{S}^{(l)}$ dans les matrices estimées $\widehat{M}^{(j,l)}_{\alpha^{(j,l)}}$ et $\widehat{M}^{(j,l)}_{II}$. Plus précisemment, ces estimateurs s'écrivent alors :

$$\widehat{M}^{(j,l)}_{\alpha^{(j,l)}} = (1-\alpha^{(j,l)})\widehat{M}^{(j,l)}_{I}\left(\widehat{\Sigma}^{(l)}\right)^{-1}\widehat{M}^{(j,l)}_{I} + \alpha^{(j,l)}\widehat{M}^{(j,l)}_{II}$$

et

$$\widehat{M}^{(j,l)}_{II} = \sum_{h=1}^{H^{(j,l)}}\frac{n^{(j,l)}_h}{n_l}\left(\widehat{V}^{(j,l)}_h - \widehat{\overline{V}}^{(j,l)}\right)\left(\widehat{\Sigma}^{(l)}\right)^{-1}\left(\widehat{V}^{(j,l)}_h - \widehat{\overline{V}}^{(j,l)}\right)'.$$

Notons par $\hat{b}$ le vecteur propre associé à la valeur propre $\hat{\lambda}$ (correspondant à la valeur propre $\lambda$ théorique). Ce vecteur $\hat{b}$ est une direction EDR estimée.

**Résultat symptotique.** Comme dans le cas homoscédastique, nous avons obtenu la covergence en probabilité à la vitesse $n^{1/2}$ de la direction EDR estimée vers la direction EDR correspondante : $\hat{b} = b + O_p(n^{-1/2})$.

# Simulations

Nous avons mis en œuvre ces deux méthodes, "PMS$_\alpha$ homoscédastique" et "PMS$_\alpha$ héteroscédastique", sur des jeux de données simulés. Les résultats des simulations sont décrits dans Liquet et Saracco (2006) et montrent la très bonne qualité des estimations obtenues.

# Références

Barreda, L., Gannoun, A. & Saracco, J. (2006). Some extensions of multivariate SIR. A paraître dans *Journal of Statistical Simulation and Computation.*

Chiaromonte, F., Cook, R.D. & Li, B. (2002) Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, **30**, 475-497.

Gannoun, A. & Saracco, J. (2003). An asymptotic theory for $SIR_\alpha$ method. *Statistica Sinica*, **13**, 297-310.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.

Liquet, B. and Saracco, J. (2006). Pooled Marginal Slicing approach via $SIR_\alpha$ with discrete covariables. A paraître dans *Computational Statistics*, **3**.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on $SIR_\alpha$. *Journal of Multivariate Analysis*, **96**, 117-135

# Gene Expression Analysis between *Vitis vinifera* and the Disease-Resistant American Grapevine *Vitis aestivalis*

**Yingcai Su**

* Adresse pour correspondance :
Department of Mathematics, Missouri State University
Springfield, MO 65804, USA
e-mail : yingcaisu@missouristate.edu

## Abstract

Microarray data analysis provides a global view of the functioning of a plant's genome. Comparing transcription profiles in different grapevines enables us to associate certain phenotypes with specific gene expression patterns which, in turn, allows us to set up hypotheses about the function of grapevine genes. Using the Affymetrix Vitis GeneChip microarray platform, we generated gene expression profiles in young leaves of *Vitis vinifera* (Cabernet sauvignon) and *Vitis aestivalis* (Norton) under normal physiological conditions. Statistical analysis of the data indicated that the Vitis GeneChip reliably measured transcript abundance in both grapevines. Genes showing differential expressions across the two species were identified, and their functions were also explored.

.

# Estimation du modèle de réduction de dimension fonctionnelle

**Anne-Françoise Yao**

* Adresse pour correspondance :
Centre d'Océanologie de Marseille,
Université de la Méditerranée, Aix-Marseille, 2.
Campus de Luminy, case 901, 13288 Marseille
e-mail : yao@com.univ-mrs.fr

## Résumé

Soit $(X, Y)$ un couple de fonctions aléatoire. On s'interesse à l'estimation de la fonction de régression : $Y = r(X) + \varepsilon$, où $\varepsilon$ est l'erreur aléatoire sous l'hypothèse du modèle de réduction de dimension : $r(x) = g(\Phi.x)$ où $g$ une fonction inconnue pouvant être estimée par une méthode à noyau et $\Phi$ un opérateur linéaire continu. Du fait de sa simplicité d'utilisation, la méthode de régression inverse fonctionnelle est d'abord utilisée pour l'estimation. Puis, dans le but de s'affranchir de l'hypothèse très contraignant d'ellipticité de la loi du régresseur, $X$ nous proposons une méthode de type ADE, utilisant la dérivée de la fonction $g(.)$.

## Références

Ferré, L. & Yao, A.F. (2005). Smoothed Functional Inverse Regression. *Statist. Sinica.* **15**, 665-683.

Hirstache, M., Juditsky, A. & Spokoiny, V. (2001). Structure adaptative approach for dimension reduction. *Ann. Statist.*, **29**, 1537-1566.

Li, K.-C. (1991). Sliced Inverse Regression for dimension reduction. *J. Amer. Statist. Assoc.* **86,** 316-342.

Ren, H. and Hsing, T. (2006). An RKHS Formulation of Inverse Regression Dimension Reduction Problem. Preprint.

Xia, Y., Härdle, W. (2006). Semi-parametric estimation of generalized partially linear single-index models. Prepint.

Xia, Y., Tong, H. and Zhu, L.X. (2002). An adaptative estimation of dimension reduction space. *J.R. Statist. Soc. B.* **64**, 363-410.