# Groupe de travail STAPH : Statistique Fonctionnelle et Opératorielle

## Partie III : Recueil de résumés 2001-2002

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

## Résumé

Ce document a pour objectif de présenter les résumés (plus ou moins détaillés selon les souhaits de leurs auteurs) des divers exposés qui ont eu lieu lors des séances du groupe de travail STAPH durant l'année universitaire 2001-2002. Rappelons que ce groupe de travail en Statistique Fonctionnelle et Opératorielle, créé il y a trois ans au sein du Laboratoire de Statistique et Probabilités de Toulouse, s'inscrit dans la dynamique actuelle autour des divers aspects fonctionnels de la statistique moderne. Les exposés qui sont présentés traitent de divers aspects de la Statistique Fonctionnelle (estimation nonparamétrique, statistique opératorielle, modèles de réduction de dimension, modèles pour variables fonctionnelles, ...) ; ils sont de nature différentes (exposés didactiques ou bibliographiques, exposés de résultats nouveaux en Statistique Appliquée et/ou Théorique, ...) ; ils témoignent enfin de l'ouverture de la démarche par la grande diversité des exposants. En préambule de ce document, un court texte est présenté afin de tirer le bilan des deux premières années de travail et afin surtout de mieux préparer l'avenir en faisant perdurer cette dynamique de recherche.

**Abstract** We present the abstracts (of size more or less important acording to the wishes of their authors) of the several different talks given during the sessions of the working group STAPH along the academic year 2001-2002. This group in Functional and Operatorial Statistics is born three years ago at the Laboratoire de Statistique et Probabilités of the Université Paul Sabatier de Toulouse, and its aim was to participate at the actual dynamic existing around the different functional features of modern statistics. These talks were about different functional topics (nonparametric estimation, statistics of operators, models for functional data, models for dimension reduction, ...). They were of different kinds (didactic, bibliografic, applied, theoric, ...) and were presented by a large variety of statisticians. As a foreword, a short text is presented to take the stock of the activities of this group during its two first years of existency in order to make an efficient preparation of the future.

**Sumario** This documento presenta resumens (mas o menos cortos segun los deseos de sus autores) de charlas que han sido presentadas durante las sesiones de trabajo del grupo STAPH durante el ano academico 2001-2002. Este grupo de trabajo en el campo de Estadistica Funcional y Operatorial ha sido creado hace tres anos en el Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier de Toulouse, para animar investigaciones en varios aspectos funcionales de la estdistica moderna. Estas conferencias fueran sobre temas variados (estimacion noparametrica, estadistica de operatores, modelos para variables funcionales, modelos de reduccion de dimension, ...) y fueran de tipos differentes (conferencias didacticas o bibliograficas, presentacion de resultados nuevos en estadistica teorica o/y applicada, ...). Al principio del documento, empezamos con un corto texto de presentacion en lo cual hacemos un chequeo de las actividades de este grupo desde dos anos y en lo cual planteamos los fundamentos para el proximo futuro.

## TABLE DES MATIERES

# [1] STAPH : Bilan de l'année 2001-2002 et perpectives

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

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

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

Pour sa troisième année d'existence le groupe de travail STAPH en Statistique Fonctionnelle et Opératorielle a poursuivi ses activités dans les trois directions essentielles que nous nous étions fixées lors de sa création :

- Statistique Multidimensionnelle et Opérateurs ;
- Estimation non-paramétrique ;
- Modèles et traitement de variables fonctionnelles.

Depuis deux ans, outre ces directions initiales, nos exposés et séances de travail se sont ouverts sur d'autres thèmes relativement connexes. Citons par exemple à cet égard :

- Choix de modèles/variables ;
- Problèmes inverses ;
- Estimation de valeurs extrêmes ;
- Analyse de Variance ;
- Déviations (petites et grandes) ;
- ...

Pour l'année qui vient de s'écouler, la diversité de nos séances reflète bien notre volonté de recherche et de mise en valeur des complémentarités entre les aspects fondamentaux de la Statistique et ses aspects plus appliqués. Dans cet esprit nous avons accueillis des exposés relatifs à divers problèmes concrets :

- Sismologie ;
- Météorologie ;
- Télécommunications ;
- ...

[1] Désormais toutes nos activités sont accessibles sur la page web

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

Par ailleurs, les sommaires de nos activités passées sont présentés à la fin de ce document.

De manière générale, cette troisième année d'existence a été placée sous le signe de l'ouverture avec une participation accrue d'intervenants extérieurs lors de nos séances de travail, mais aussi lors des Journées de Statistique Fonctionnelle que nous avons organisées à Toulouse les 10 et 11 Juin 2002 en partenariat avec l'Université de Sidi-bel-Abbes et l'INRA de Toulouse.

Un grand merci donc à tous les intervenants et tous les participants à nos séances de travail, ainsi qu'à tous ceux qui ont manifesté l'intérêt qu'ils portaient à notre initiative. Tout ceci ne peut que nous pousser à poursuivre dans cette direction.

Merci à tous et bonnes vacances.[2]

[2]Désormais toutes nos activités sont accessibles sur la page web

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

Par ailleurs, les sommaires de nos activités passées sont présentés à la fin de ce document.

# Estimation fonctionnelle d'un operateur de transition d'un processus de Markov à états continus

**Abderrahmane YOUSFATE**

Université de Sidi Bel Abbès. Algérie
Laboratoire de Mathématique
BP 89 Sidi Bel Abbès. 22 000. Algérie
e-mail : yousfate_a@yahoo.com

Exposé du 16 Octobre 2001

## Résumé

Dès le début du vingtième siecle, les chaîne de Markov ont été exploitées comme outil de prévision par Markov sans que le formalisme mathématique (voir Kolmogorov) des probabilités et de la statistique mathématique ne soit établi. Il s'est avéré par la suite que les techniques utilisées correspondaient à celles du maximum de vraissemblance. Albert A. (1962) a construit des estimateurs à temps continu et à états discrets pour le générateur du processus dont les résultats servent à estimer l'opérateur de transition. Pour ce qui est des processus de Markov à états continus (temps discret ou continu) une littérature importante a été écrite dans ce domaine (voir Doob (1953), Gihman et Skorohod (1970-1979) par exemple) pour décrire les structures de ces processus.

Pour faire une estimation de la densité de l'opérateur de transition d'un processus de Markov à temps discret et à états continus, Roussas (1969) a été le premier à utiliser l'estimation fonctionnelle. Parmi les références générales dans ce domaine citons Collomb et Doukhan (1983), Bosq (1996), Roussas (1991) et Ferraty et Vieu (2001). Des articles plus spécifiquement consacrés à l'étude de densité conditionnelle sont ceux de Youndjé (1996) et Hyndman et Yao (2001).

Sur un autre plan, beaucoup d'auteurs (par exemple Karlin : 1966 et 1981, Iosifescu et Tăutu, 1980, Yousfate, 1986, ...) ont utilisé des propriétés matricielles de la transition pour y associer des propriétés structurelles des chaînes de Markov étudiées.

En tenant compte de certaines propriétés spectrales de l'opérateur de transition d'un processus de Markov, un estimateur fonctionnel sera étudié selon certaines hypothèses privilégiant l'échange entre états du processus. Une majoration de la vitesse de convergence (au sens $L^p$) sera présentée. Les résultats que nous présentons sont une extension de ceux tirés de Laksaci et Yousfate (2001).

# Références

Albert, A. (1962) Estimating the infinitesimal generator of continuous time fifnite state Markov process, *Annals of Mathematical Statistics*, **33**, no. 2, p 727-753.
Bosq, D. (1996) *Nonparametric statistics for stochastic processes*, Lecture Note in Statitics, **110**, Springer Verlag. Berlin.
Collomb, G. et Doukhan, P. (1983) Estimation non paramétrique de la fonction d'autoregression d'un processus stationnaire $\varphi$-mélangeant : risques quadratiques par la méthode du noyau, *CRAS*, **296**, Série I, p 859-862.
Doob, J.L. (1953) *Stochastic processes*. Wiley. New York.
Ferraty, F. et Vieu P. (2001) *Statistique Fonctionnelle : modèles de régression pour variables aléatoires uni, multi et infiniment dimensionnées*, Publication du Laboratoire de Statistique et probabilités de Toulouse, **LSP 2001-03**.
Gihman, L.I. et Skorohod, A.V. (1970 Tome I)(1975 Tome II)(1979 Tome III) *Theory of stochastic processes.*, Springer Verlag. Berlin
Hyndman, R.J.et Yao, Y. (2001) *Nonparametric estimation and symetry tests for conditional density functions. Preprint.*
Iosifescu, M. et Tăutu P. (1980) *Finite Markov processes and their applications.* Wiley. New York.
Karlin, S. (1966) *A first course in stochastic processes*, Academic Press, New York.
Karlin, S. (1981) *A second course in stochastic processes*, Academic Press, New York.
Laksaci, A. et Yousfate A. (2001) Estimation fonctionnelle de la densité de l'opérateur de transition d'un processus de Markov à temps discret. *Preprint.*
Roussas, G. (1969) Nonparametric estimation of the transition distribution function of a Markov process. *Annals of mathematical statistics*, **40**, p 1386-1400.
Roussas, G. (1991) Recursive estimation of the transition distribution function of a Markov process. Asymptotic normality. *Statistics and Probability Letters*, **11**, p 435-447.
Youndjé, E. (1996) Propriétés de convergence de l'estimateur à noyau de la densité conditionnelle. *Rev. Roumaine de math. pures et appl.*, **41**, 7-8, p 535-566.
Yousfate, A. (1986) Décomposition canonique d'un processus qualitatif de type markovien stationnaire. *Statistique et Analyse des données*, **11**, p 64-89.

# PLS regression on a stochastic process

**Christian PREDA ***
En collaboration avec Gilbert SAPORTA

* Adresse pour correspondance :
Département de Statistique, Faculté de Médecine
Université de Lille 2, 1, Place de Verdun, 59045 Lille Cedex, France
e-mail : cpreda@univ-lille2.fr

Exposé du 12 Novembre 2001

## Abstract

We give an extension of PLS regression to the case where the set of predictor variables forms a $L_2$-continuous stochastic process and the response is a random vector of finite or infinite dimension. We prove the existence of PLS components as eigenvectors of some operator and also some convergence properties of the PLS approximation. The results of an application to stock-exchange data will be compared with those obtained by others methods.

**Keywords.** PLS regression, Escoufier's operator, principal component analysis.

## 1. Introduction

It doesn't seem usual to perform a linear regression when the number of predictors is infinite. However it is the case when one tries to predict a response variable $Y$ thanks to the observation of a time dependent variable $X_t$, for any $t \in [0, T]$ (for example, $(X_t)_{t\in[0,T]}$ can represent temperature curves observed in $n$ places and $Y$ the amount of crops). Theoretically, this can be expressed by the regression of the $Y$ variable on the $(X_t)_{t\in[0,T]}$ process. he aim of this article is to adapt the PLS regression when all the explicative variables form a stochastic process. The problems brought about by the classical linear regression on a process – the indetermination of the regression coefficients (Ramsay and Dalzell (1991), Ramsay and Silverman (1997), Saporta (1981)) or the choice of the principal components of $(X_t)_{t\in[0,T]}$ as explicative variables (Deville (1978),

Saporta (1981), Aguilera et al. (1998)) – get within this framework satisfactory solutions the main characteristics of which derive from those of the Escoufier operator associated with the process $(X_t)_{t\in[0,T]}$ (Saporta (1981)).

PLS regression on a stochastic process is an extension of the finite case (a finite set of predictors) developed by Wold et al.(1984), Tenenhaus et al. (1995) and Cazes (1997). We prove the existence of PLS components as well as a few convergence properties towards the classical linear regression. The case $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$, presents an alternative to prevision methods proposed by Aguilera et al. (1998) and Deville (1978). The results of an application on stock exchange data are compared with those obtained by other methods.

# 2. Main results

It doesn't seem usual to perform a linear regression when the number of predictors is infinite. However it is the case when one tries to predict a response variable $Y$ thanks to the observation of a time dependent variable $X_t$, for any $t \in [0,T]$ (for example, $(X_t)_{t\in[0,T]}$ can represent temperature curves observed in $n$ places and $Y$ the amount of crops). Theoretically, this can be expressed by the regression of the $Y$ Let $(X_t)_{t\in[0,T]}$ be a random process and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, $p \geq 1$, a random vector defined on the same probability space $(\Omega, \mathcal{A}, P)$. We assume that $(X_t)_{t\in[0,T]}$ and $\mathbf{Y}$ are of second order, $(X_t)_{t\in[0,T]}$ is $L_2$-continuous and for any $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is an element of $L_2([0,T]$. Without loss of generality we assume also that $E(X_t) = 0$, $\forall t \in [0,T]$ and $E(Y_i) = 0$, $\forall i = 1, \ldots, p$.

Under this hypothesis, let $\mathbf{C}_{YX}$ and $\mathbf{C}_{XY}$ be the operators defined as :

$$\mathbf{C}_{YX} : L_2([0,T]) \to \mathbf{R}^p,\ f \xrightarrow{\mathbf{C}_{YX}} x, \quad x_i = \int_0^T E(X_t Y_i) f(t) dt, \quad \forall i = 1, ..., p,$$

$$\mathbf{C}_{XY} : \mathbf{R}^p \to L_2([0,T]),\ x \xrightarrow{\mathbf{C}_{XY}} f, \quad f(t) = \sum_{i=1}^{p} E(X_t Y_i) x_i, \quad \forall t \in [0,T].$$

and denote by $\mathbf{U}_X = \mathbf{C}_{XY} \circ \mathbf{C}_{YX}$ and by $\mathbf{U}_Y = \mathbf{C}_{YX} \circ \mathbf{C}_{XY}$.

The following proposition justifies these definitions and gives the solution to the PLS problem :

**Proposition 1 (Tucker criterion)**

$$\begin{array}{rl} \max\limits_{w,\,c} & Cov^2\left(\int_0^T X_t w(t) dt, \sum_{i=1}^{p} c_i Y_i\right) \\ w \in L_2([0,T]), \|w\| = 1 & \\ c \in \mathbf{R}^p, \|c\| = 1 & \end{array}$$

*is reached for $w$, respectively $c$, the eigenvectors associated to the largest eigenvalue of $\mathbf{U}_X$, respectively of $\mathbf{U}_Y$.*

Let $w_1 \in L_2([0,T])$ be the eigenfunction of $\mathbf{U_X}$ associated to the largest eigenvalue. Then, the first PLS component (Tenenhaus et al. (1995)) of the regression of $\mathbf{Y}$ on the process $(X_t)_{t\in[0,T]}$ is the random variable defined as :

$$t_1 = \int_0^T X_t w_1(t)dt \tag{1}$$

Denote by $\mathbf{W}^X$, respectively $\mathbf{W}^Y$, the Escoufier's operators[3] (Escoufier (1970)) associated to $(X_t)_{t\in[0,T]}$, respectively to $\mathbf{Y}$, defined by :

$$\mathbf{W}^X Z = \int_0^T E(X_t Z)X_t dt, \quad \mathbf{W}^Y Z = \sum_{i=1}^p E(Y_i Z)Y_i, \quad \forall Z \in L_2(\Omega).$$

Our main result is the following theorem.

**Theorem 2** *$t_1$ is the eigenvector of the $\mathbf{W}^X\mathbf{W}^Y$ associated to the largest eigenvalue.*

The PLS regression is an iterative method. Let $X_{0,t} = X_t$, $\forall t \in [0,T]$ and $Y_{0,i} = Y_i$, $\forall i = 1,...,p$. At the step $h$, $h \geq 1$, of the PLS regression of $\mathbf{Y}$ on $(X_t)_{t\in[0,T]}$, we define the $h^{\text{th}}$ PLS component, $t_h$, by the eigenvector associated to the largest eigenvalue of the operator $\mathbf{W}^X_{h-1}\mathbf{W}^Y_{h-1}$,

$$\mathbf{W}^X_{h-1}\mathbf{W}^Y_{h-1}t_h = \lambda_{\max}t_h, \tag{2}$$

where $\mathbf{W}^X_{h-1}$, respectively $\mathbf{W}^Y_{h-1}$, are the Escoufier's operators associated to $(X_{h-1,t})_{t\in[0,T]}$, respectively to $\mathbf{Y}_{h-1} = (Y_{h-1,i})_{i=1,...,p}$ and $\lambda_{\max}$ the largest eigenvalue of $\mathbf{W}^X_{h-1}\mathbf{W}^Y_{h-1}$. Finally, the PLS step is completed by the ordinary linear regression of $X_{h-1,t}$ and $Y_{h-1,i}$ on $t_h$. Let $X_{h,t}$, $t \in [0,T]$ and $Y_{h,i}$, $i = 1,...,p$ be the random variables which represent the error of these regressions :

$$X_{h,t} = X_{h-1,t} - p_h(t)t_h, \quad t \in [0,T],$$

$$Y_{h,i} = Y_{h-1,i} - c_{h,i}t_h, \qquad i = 1,...,p,$$

As in the finite case (Tenenhaus et al. (1995)), the next statements hold :

**Proposition 3** *For each $h \geq 1$ :*
*a) $\{t_h\}_{h\geq 1}$ forms an orthogonal system in $L_2(X)$,*
*b) $Y_i = c_{1,i}t_1 + c_{2,i}t_2 + \ldots + c_{h,i}t_h + Y_{h,i}, \quad i = 1,...,p,$*

[3]The spectral analysis of this operator leads to the principal component analysis of the associated variable. See Deville (1974) and Escoufier (1970) for details.

*c)* $X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_h(t)t_h + X_{h,t}, \quad t \in [0,T],$
*d)* $E(Y_{h,i}t_j) = 0, \quad \forall i = 1,...,p, \forall j = 1,...,h,$
*e)* $E(X_{h,t}t_j) = 0, \quad \forall t \in [0,T], \forall j = 1,...,h.$

From the Proposition 3-b), the PLS approximation of $\mathbf{Y}$ by $(X_t)_{t\in[0,T]}$ at step $h$, $h \geq 1$, is given by :

$$\hat{\mathbf{Y}}_h = c_1t_1 + \ldots + c_ht_h, \quad c_i \in \mathbf{R}^p, i = 1,...,p. \tag{3}$$

Denote by $\hat{\mathbf{Y}}$ the approximation of $\mathbf{Y}$ given by the ordinary linear regression on $(X_t)_{t\in[0,T]}$. Then, the sequence $\{\hat{\mathbf{Y}}_h\}_{h\geq 1}$ is convergent in $L_2(\Omega)$ and the limit is $\hat{\mathbf{Y}}$ :

**Proposition 4**

$$\lim_{h\to\infty} E(\|\hat{\mathbf{Y}}_h - \hat{\mathbf{Y}}\|^2) = 0 \tag{4}$$

Finally, the choice of $h$ using the cross-validation criterion (Green and Silverman (1994), Tenenhaus (1998)) remains applicable in this case.

**Remark 5**

a) **(The continuous case)** The previous results are still valid for the particular case $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$. Indeed, because of the $L_2$ continuity of the process $(X_t)_{t\in[0,T+a]}$, $\mathbf{C}_{X,Y}$ and $\mathbf{C}_{Y,X}$ are compacts and therefore, $\mathbf{U}_X$ and $\mathbf{U}_Y$ are compacts. The results of the Proposition 1 and Theorem 2 are preserved.
The decomposition formulas (Proposition 3-b,c) become in this case :

$$X_t = \begin{cases} t_1p_1(t) + \ldots t_hp_h(t) + X_{h,t}, & \forall t \in [0,T], \\ \\ t_1c_1(t) + \ldots t_hc_h(t) + X_{h,t}, & \forall t \in [T,T+a], \end{cases} \tag{5}.$$

For each $s \in [0,a]$, the "forecast" of $X_{T+s}$ by $(X_t)_{t\in[0,T]}$ is given by :

$$\hat{X}_{T+s} = t_1c_1(T+s) + \ldots + t_hc_h(T+s). \tag{6}$$

b) **(Approximation)** Let $\Delta = \{0 = t_0 < t_1 < \ldots < t_p = T\}$, $p \geq 1$, be a discretization of $[0,T]$ and consider the process $(X_t^\Delta)_{t\in[0,T]}$ defined as (Preda (1999)) :

$$X_t^\Delta = \frac{1}{t_{i+1} - t_i} \int_{t_i}^{t_{i+1}} X_t dt, \quad \forall t \in [t_i, t_{i+1}[, \quad \forall i = 0,...,p-1.$$

Denote by $m_i$ the random variable $\frac{1}{t_{i+1}-t_i}\int_{t_i}^{t_{i+1}} X_t dt$, $i = 0,\ldots,p-1$. Then, the approximation of the PLS regression on $(X_t)_{t\in[0,T]}$ by those on $(X_t^\Delta)_{t\in[0,T]}$

is equivalent to the PLS regression on the finite set $\{m_i\sqrt{t_{i+1}-t_i}\}$, $i = 0,\ldots,p-1$.

# Application on stock exchange data

The PLS regression on a process presented in the previous sections will be used to predict the behaviour of shares on a certain lapse of time.

We have 84 shares quoted at the Paris stock exchange, for which we know the whole behavior of the growth index during one hour (between $10^{00}$ and $11^{00}$); a share is likely to change every second. We also know the evolution of the growth index of a new share (noted 85) between $10^{00}$ and $10^{55}$. The aim is to predict the way that share will behave between $10^{55}$ and $11^{00}$ using a PLS model built with the other 84 shares.

Using the SIMCA-P software (see Tenenhaus (1998) for details) we are going to build several PLS models according to the number of components chosen for the regression. So we are going to refer to the model with $k$ PLS components as PLS(k). The previsions obtained with those models will be compared to those given by the regression on the principal components (models quoted with CP(k)) and the algorithm NIPALS (see Tenenhaus (1998) for details). To rate the quality of those models we are going to compare the previsions obtained by each model, quoted with $\{\hat{m}\}_i(85)$, with true values $\{m\}_i(85)$, $i = 56,\ldots,60$, observed previously. The forecasts of these models are presented in the next Table.

| | $\hat{m}_{56}(85)$ | $\hat{m}_{57}(85)$ | $\hat{m}_{58}(85)$ | $\hat{m}_{59}(85)$ | $\hat{m}_{60}(85)$ | $SSE = \sum_{i=56}^{60} (\hat{m}_i - m_i)^2$ |
|---|---|---|---|---|---|---|
| **Observed** | **0.700** | **0.678** | **0.659** | **0.516** | **-0.233** | **-** |
| PLS(1) | -0.327 | -0.335 | -0.338 | -0.325 | -0.302 | 3.789 |
| PLS(2) | 0.312 | 0.355 | 0.377 | 0.456 | 0.534 | 0.928 |
| PLS(3) | 0.620 | 0.637 | 0.677 | 0.781 | 0.880 | 1.318 |
| CP(1) | -0.356 | -0.365 | -0.368 | -0.355 | -0.331 | 4.026 |
| CP(2) | -0.332 | -0.333 | -0.335 | -0.332 | -0.298 | 3.786 |
| CP(3) | 0.613 | 0.638 | 0.669 | 0.825 | 0.963 | 1.538 |
| NIPALS | 0.222 | 0.209 | 0.240 | 0.293 | 0.338 | 1.000 |

# References

AGUILERA, A.M., OCAÑA F. and VALDERAMA, M.J. (1998) : *An approximated principal component prediction model for continous-time stochastic process*, Applied Stochastic Models and Data Analysis, Vol. 13, p. 61-72.

CAZES, P. (1997) : *Adaptation de la régression PLS au cas de la régression après Analyse des Correspondances Multiples*, Revue de Statistique Appliquée, XLIV (4), p. 35-60.
DEVILLE, J.C. (1974) : *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE, No. 15, p 3-101.
DEVILLE, J. C. (1978) : *Analyse et prévision des séries chronologiques multiples non stationnaires*, Statistique et Analyse des Données, No. 3, p. 19-29.
ESCOUFIER, Y. (1970) : *Echantillonnage dans une population de variables aléatoires réelles*, Publications de l'Institut de Statistique de l'Université de Paris, 19, Fasc. 4, p. 1-47.
GREEN, P.J. and SILVERMAN, B. W. (1994) : *Nonparametric Regression and generalized linear models. A roughness penalty approach*, Monographs on statistic and applied probability, No. 58, Chapman & Hall.
LEBART, L., MORINEAU, A. and PIRON, M. (1995) : *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
PALM, R. and IEMMA, A.F. (1995) : *Quelques alternatives à la régression classique dans le cas de colinéarité*, Rev. Statistique Appliquée XLIII (2), p. 5-33.
PREDA, C. (1999) : *Analyse factorielle d'un processus : problèmes d'approximation et de régression* , Thèse de doctorat de l'Université de Lille 1, No. 2648.
RAMSAY, J.O. and DALZELL, C.J. (1991) : *Some tools for functional data analysis*, Journal of Royal Statistical Society (B), 53, No. 3, p. 539-572.
RAMSAY, J.O. and SILVERMAN, B.W. (1997) : *Functional Data Analysis*, Springer Series in Statistics, Springer-Verlag, New York.
SAPORTA, G. (1981) : *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O., No. 37-38, Université Pierre et Marie Curie, Paris.
TENENHAUS, M., GAUCHI, J.P. and MENARDO, C. (1995) : *Régression PLS et applications*, Revue de Statistique Appliquée, XLIII (1), p. 7-63.
TENENHAUS, M. (1998) : *La régression PLS. Théorie et pratique*, Editions Technip, Paris.
WOLD S., RUHE A., DUNN W.J. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput., vol 5, no.3 pp. 735-743.

# Réconcilions ridge regression et troncature spectrale en testant la moyenne d'une courbe aléatoire.

**André MAS**

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

Exposé du 26 Novembre 2001

---

## Résumé

En dimension finie, le test d'hypothèse nulle $m = m_0$ où m est la moyenne de l'échantillon est obtenu en inversant (quand cela est possible) la matrice de covariance empirique. Dans le cas de v.a. infini-dimensionnelles, une approche similaire mène à une impasse. On propose une statistique de test basée sur l'inversion par pénalisation de l'opérateur de covariance empirique des courbes aléatoires. Cette étude peut être ramenée en termes simples à un probleme linéaire inverse mal posé. L'originalité de l'approche consiste à lier de façon sous-jacente le paramètre de pénalisation à une troncature spectrale.

# Estimation fonctionnelle et Ondelettes

**Antoine AYACHE et Jean Michel LOUBES**

Université Paul Sabatier
Laboratoire de Statistique et Probabilités
Toulouse
ayache@cict.fr
loubes@cict.fr, loubes@dptmaths.ens-cachan.fr

Exposé du 10 Décembre 2001

---

## 1. Introduction aux ondelettes

Dans cette partie nous nous proposons d'introduire certains concepts fondamentaux de la théorie des ondelettes. *Une base d'ondelettes* orthonormales (dyadiques) $\mathcal{M}$ de l'espace de Hilbert $L^2(\mathbb{R}^d)$ est une base générée par dilatation et par translation de $2^d - 1$ fonctions $\psi_i$, plus précisément

$$\mathcal{M} = \{2^{jd/2}\psi_i(2^j x - k),\ i \in \{1, \dots 2^d - 1\},\ j, k \in Z\}.$$

Haar, dans sa thèse sous la direction de Hilbert, avait déjà introduit en 1909, une base de ce type. *La notion d'analyse multirésolution* (AMR) founit un procédé de construction de bases d'ondelettes, qui est à la fois naturel et simple. Cette notion a été mise au point par Mallat et Meyer en 1986. Il s'agit de la donnée d'une suite de sous-espaces $\{V_j\}_{j\in Z}$ de $L^2(\mathbb{R}^d)$, fermés, emboités et vérifiant les propriétés suivantes :

(a) $\cap_{j=-\infty}^{+\infty} V_j = \{0\}$ et $\overline{\cup_{j=-\infty}^{+\infty} V_{j\in Z}} = L^2(\mathbb{R}^d)$,
(b) $\forall\ j \in Z,\ f(x) \in V_j \Longleftrightarrow f(2x) \in V_{j+1}$,
(c) $V_0 = \overline{Vect\{\varphi(x-k),\ k \in Z^d\}}$.

On dit que $\varphi$ est une fonction d'échelle, parce que sa transformée de Fourier vérifie une relation du type $\hat{\varphi}(2\xi) = m_0(\xi)\hat{\varphi}(\xi)$, où $m_0(\xi)$ est un Filtre en Quadrature Conjugué (CQF). Ainsi, il existe une correspondance entre les AMR et les CQF's. C'est d'ailleurs en partant de CQF's ad hoc que Daubechies a pu construire une famille de bases d'ondelettes à support compact, très utiles dans certaines applications.

# 2. Application des ondelettes à la théorie de l'estimation

Dans cette partie, nous nous proposons d'appliquer la théorie des ondelettes à l'estimation non paramétrique. En effet, les ondelettes font parties des meilleures bases pour représenter des objets présentant des singularités dont on ne connaît ni le nombre, ni laposition.

Dans un premier temps, nous présenterons des estimateurs obtenus par projection sur une base d'ondelettes ainsi que les estimateurs de seuillage, étudiés par Donoho, Johnstone, Kerkyacharyan et Picard (1997-1998). Nous étudierons leur performance asymptotique et monterons que de tels estimateurs sont adaptatifs, c'est-à-dire qu'ils atteignent la vitesse minimax de convergence tout en étant définis sans connaîte a priori la régularité du problème.

Dans un second temps, nous monterons comment les bases d'ondelettes peuvent servir à modéliser les fonctions multifractales (Jaffard 2001) et comment il est possible d'estimer de telles fonctions (Gamboa et Loubes 2001).

# Références

Aubry and S. Jaffard, (2001). Random wavelet series. *Technical Report.*
I. Daubechies, (1992). *Ten lectures on wavelets.* SIAM Philadelphia, CBMS-NSF series, volume **61**.
D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, (1995). Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, **57(2)**, 301-369, with discussions.
D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, **24(2)**, 508-539.
S. Jaffard, Y. Meyer, and R.D. Ryan, (2001). *Wavelets. Tools for science & technology.*, SIAM, (Update of Y. Meyer : Wavelets : Algorithms & Applications).
S. Mallat, (1998). *A wavelet tour of signal processing.*, Academic Press.
Y. Meyer, (1990). *Ondelettes et opérateurs I : ondelettes.* Actualités Mathématiques, Herman, Paris.
Y. Meyer, (1992). *Ondelettes et algorithmes concurrents.* Hermann, Paris.
H.L. Resnikoff and R.O. Wells Jr, (1998). *Wavelet Analysis : The Scalable Structure of Information..* Springer, New York.

Pour une bibliographie plus complète voir :
*http ://www.cs.kuleuven.ac.be/ ade/WWW/WAVE/references.html*

# Analyse de l'activité d'un centre de renseignement téléphonique : étude par modèle additif avec composante d'interaction de dimension réduite

**Simplice DOSSOU-GBÉTÉ**

Laboratoire de Mathématiques Appliquées
Université de Pau et des Pays de l'Adour
e-mail : simplice.dossou-gbete@univ-pau.fr

Exposé du 17 Décembre 2001

---

**Mots clés** : modèle poissonnien, modèle linéaire généralisé, modèle additif, données fonctionnelles, vraisemblance locale, moindres carrés asymptotiques, interaction de dimension réduite.

## Introduction

Ce travail propose l'utilisation de méthodes d'analyse de données fonctionnelles (Ramsay J.O. & Silverman B.W., 1997) pour étudier les variations hebdomadaires et journalières du nombre des appels servis (i.e. appels ayant reçu une réponse avant abandon par le client) au centre régionnal des renseignements téléphoniques de la société France Télécom situé à Pau (France). Les appels servis chaque semaine sont ventilés dans des tableaux suivant les sept jours de la semaine et 29 tranches horaires consécutives et adjacentes de 30 minutes chacune. Le modèle statistique qui soustend cette étude est basé sur la famille des lois de probabilité de Poisson. Comme dans Green & Silverman (1994), la modèlisation envisagée est une extension naturelle du modèle linéaire généralisé de McCullagh & Nelder (1989) aux données fonctionnelles. Le modèle proposé dans cet article considère que le nombre moyen des appels servis résulte d'une combinaison d'effets spécifiques du jour de la semaine et de la tranche horaire ainsi que d'une interaction jour-tranche horaire. Les effets spécifiques des tranches horaires ainsi que les interactions "jour-tranche horaire" sont modélisés par des paramètres fonctionnels. On considère de plus que les paramètres fonctionnels qui tranduisent les interactions appartiennent à un espace vectoriel de dimension réduite. Cet type de modélisation est très utilisé pour l'analyse des tables de contingence (Godmann, Baccini, Caussinus et de Falguerolles, 1993).

Dans cette étude, nous proposons l'utilisation de la méthode des moindres

carrés asymptotiques dans un contexte d'analyse de données fonctionnelles. L'utilisation de la méthode des moindres carrés pour l'estimation des paramètres d'un modèle de régression linéaire généralisé est assez ancienne. Taylor (1953) l'étudie dans le cadre de l'estimation des paramètres d'un modèle de régression logistique (Méthode de Berkson et méthode du Khi-Deux minimum). Plus récemment, Baccini, Caussinus et de Falguerolles (1993) ont repris les arguments de Taylor pour justifier l'utilisation des moindres carrés comme alternative au maximum de vraisemblance pour l'estimation des paramètres dans les modèles d'association utilisés pour l'analyse des tables de contingence. Les estimateurs proposés par Taylor et comme ceux de Baccini, Caussinus et de Falguerolles ne constituent en fait que des versions spécialisées de la méthode des moindres carrés asymptotiques exposée en particulier dans Gourriéroux et Monfort (1989). Cette méthode est exposée aussi dans Dobson (1983) comme la Delta-Méthode. Dans le cadre du modèle linéaire généralisé, la méthode des moindres carrés asymptotiques fournit en effet des estimateurs convergents et asymptotiquement gaussiens (Dobson, 1983, Gouriéroux et Monfort, 1989).

# Données

Ce travail a été motivée par des données qui nous ont été communiquées par le service régional des renseignements téléphoniques de France Télécom basé à Pau. Tout appel arrivé à ce centre des renseignements téléphoniques entre 07h30 et 22h est enregistré par l'ordinateur du service dans une base de données. Certains appels sont abandonnés avant que le service n'ait pu leur donner une réponse. Les raisons de ces abandons ne sont pas encore identifées. Un délai de réponse long ainsi qu'une erreur de numérotation de la part du client sont des causes d'abandon possibles. Les données analysées dans cette étude sont les nombres d'appels servis (c'est-à-dire les appels ayant obtenu une réponse) par période consécutive d'une demi-heure entre 07h30 et 22h, du lundi au dimanche, pendant les semaines 38, 40 et 41 de l'année 1998.
Soit $\{1, 2, ...7\}$ l'énumération chronologique des jours de la semaine et $\{1, 2, ...29\}$ celle des différentes tranches horaires d'une demi-heure de 07h30 à 22h. On note alors $n_{sjh}$ le dénombrement des appels servis pendant la tranche horaire de rang h, le jour j de la semaine $s = 1, 2, 3$ . Il semble raisonnable de considérer que les appels arrivent au centre des renseignements téléphoniques indépendamment les uns des autres. En conséquence, si on tient compte des conclusions de l'analyse exploratoire des données effectuée à la section précédente, on peut considérer que $n_{sjh}$ est la réalisation d'une variable aléatoire de Poisson $N_{sjh}$ et que

$$\{N_{sjh},\ s = 1, 2, 3,\ j \in \{1, \ldots, 7\},\ h \in \{1, \ldots, 29\}\}$$

est une suite de variables aléatoires indépendantes.

# Modèles statistiques

-1. Modèle additif fonctionnel saturé.

Il vient de l'examen des données que $E(N_{sjh})$ dépend seulement du jour j, de la tranche horaire h et d'une interaction entre le jour et la tranche horaire. On pose alors $\lambda_{jh} = E(N_{sjh})$ . Nous proposons de modéliser les caractéristiques des variations du nombre moyen d'appels servis sous forme additive à l'aide de la relation suivante

$$\theta_{jh} = g(\lambda_{jh}) = \alpha_j + \beta_h + \gamma_{jh}$$

où :

($i$) g est un fonction de lien (i.e. une fonction numérique bijective et deux fois différentiable définie sur le domaine de variation des paramètres $\lambda_{jh}$)
($ii$) $\alpha_j$ exprime l'effet spécifique du jour j,
($iii$) $\beta_h$ exprime l'effet spécifique de la tranche horaire h d'une journée
($iv$) $\gamma_{jh}$ traduit l'interaction entre le jour j et la période horaire h

On suppose que $\beta_h$ et $\gamma_{jh}$ sont respectivement les évaluations en $\frac{h-1}{28}$ de fonctions numériques $\beta$ et $\gamma_j$ définies sur l'intervalle $[0,1]$ et appartenant à un espace de Hilbert $F$ de fonctions numériques régulières définies sur $[0,1]$. Il résulte de ces hypothèses que $\theta_{jh}$ et $\lambda_{jh}$ sont, pour tout j, les évaluations en $\frac{h-1}{28}$ de fonctions régulières $\theta_j$ et $\lambda_j$ définies sur $[0,1]$. Les paramètres du modèle sont : la suite numérique $\left\{\alpha_j, j = \overline{1,7}\right\}$, les fonctions $\beta$ et $\gamma_j$ $\left(j = \overline{1,7}\right)$. Lorsqu'il est identifiable, un tel modèle est dit saturé.

L'intérêt de considérer des paramètres fonctionnels $\beta$ et $\gamma_j$ tient au fait qu'il permet de prendre en compte les situations où les suites $\left\{\lambda_{jh},\ h = \overline{1,H_j}\right\}$ ne sont pas de même longueur H.

-2. Modèle additif fonctionnel avec interaction de dimension réduite.

Le modèle additif fonctionnel avec interaction de rang réduit se définit à partir du modèle additif précédent avec l'hypothèse additionnelle que les fonctions $\gamma_j$ appartiennent à un même sous-espace vectoriel $U$ de dimension q dans $F$. Les paramètres du modèle sont alors : la suite numérique $\left\{\alpha_j, j = \overline{1,7}\right\}$, les fonctions $\beta$ et $\gamma_j$ $\left(j = \overline{1,7}\right)$, le sous-espace vectoriel $U$ et sa dimension q.

-3. Identification du modèle additif fonctionnel.

Le modèle ci-dessus n'est pas identifiable dans la mesure où il n'y a pas unicité des paramètres $\alpha_j, j = \overline{1,7}$, $\beta$ et $\gamma_j$, $j = \overline{1,7}$. On peut assurer l'identifiabilité du modèle par des contraintes additionnelles sur ses paramètres. Ces contraintes

d'identifiabilité s'expriment généralement sous la forme

$$\sum_{j=1}^{7} \pi_j \alpha_j = \alpha_0, \sum_{j=1}^{7} \pi_j \gamma_j = 0,$$

$$\int_0^1 \tau(x)\,\beta(x)\,dx = \beta_0, \int_0^1 \tau(x)\,\gamma_j(x)\,dx = 0.$$

où $\left\{\pi_j, j = \overline{1,7}\right\}$ est une suite numérique strictement positive et $\tau$ est une fonction positive définie sur $[0,1]$. On peut supposer, sans perte de généralité, que $\tau$ est une densité de probabilité sur $[0,1]$ et que $\sum_{j=1}^{7} \pi_j = 1$. L'extension au cas des données fonctionnelles de la pratique courante dans les modèles d'association pour l'analyse des tables de contingence nous amène à envisager le choix ci-dessous

$$\pi_j = \frac{\int_0^1 \lambda_j(x)\,dx}{\overline{\lambda}}, \tau = \frac{\sum_{j=1}^{7} \lambda_j}{\overline{\lambda}}, \alpha_0 = \beta_0 = \frac{1}{2}\overline{\lambda}$$

avec $\overline{\lambda} = \sum_{j=1}^{7} \pi_j \int_0^1 \lambda_j(x)\,dx$

# Estimation des paramètres

-1. Estimation des paramètres du modèle saturé.

Dans le cas du modèle saturé, il résulte des contraintes d'identifiabilité que

$$\alpha_j = \int_0^1 \tau(x)\,\theta_j(x)\,dx - \beta_0, \beta = \sum_{j=1}^{7} \pi_j \theta_j - \alpha_0, \gamma_j = \theta_j - \alpha_j - \beta.$$

Si :

($a$) $\widehat{\theta}_j$ et $\widehat{\tau}$ sont respectivement des estimateurs fonctionnels consistants (par exemple des estimateurs du maximum de vraisemblance) de $\theta_j$ et $\tau$,

($b$) $\widehat{\pi}_j$, $\widehat{\beta}_0$ et $\widehat{\alpha}_0$ sont des estimateurs consistants de $\pi_j$, $\beta_0$ et $\alpha_0$ respectivement,

alors

$$\begin{aligned}
\widehat{\alpha}_j &= \int_0^1 \widehat{\tau}(x)\,\widehat{\theta}_j(x)\,dx - \widehat{\beta}_0 \\
\widehat{\beta} &= \sum_{j=1}^{7} \widehat{\pi}_j \widehat{\theta}_j - \widehat{\alpha}_0 \\
\widehat{\gamma}_j &= \widehat{\theta}_j - \widehat{\alpha}_j - \widehat{\beta}
\end{aligned}$$

sont des estimateurs consistants de $\alpha_j$, $\beta$ et $\gamma_j$ et asymptotiquement identifiables.

Dans le cas du modèle additif paramétrique avec interaction de rang réduit, la méthode des moindres carrés asymptotiques permet d'envisager des estimateurs consistants des $\alpha_j$, $\beta_h$, $\gamma_{jh}$ et $U$ dès lors que l'on dispose d'estimateurs presque-sûrement convergents de $\tau_h$ , $\theta_{jh}$, $\pi_j$, $\beta_0$ et $\alpha_0$ (Gouriéroux Ch. & Monfort A., 1989, pp.301-314). Nous proposons d'étendre cette méthode au cas des données fonctionnelles de la manière décrite dans la section qui va suivre.

-2. Méthode des moindres carrés asymptotiques pour données fonctionnelles.

Soient $\widehat{\pi}_j$, $\widehat{\tau}$, $\widehat{\theta}_j$, $\widehat{\alpha}_0$ et $\widehat{\beta}_0$ des estimateurs fonctionnels de $\pi_j$, $\tau$, $\theta_j$, $\alpha_0$ et $\beta_0$. On pose pour tout $f$, $f_1$ et $f_2$ dans $F$

$$\begin{aligned}(f_1 \mid f_2)_{\widehat{\tau}} &= \int_0^1 \widehat{\tau}(x) f_1(x) f_2(x)\, dx \\ \|f\|^2_{\widehat{\tau}} &= \int_0^1 \widehat{\tau}(x) f^2(x)\, dx\end{aligned}$$

Soit $vect(F,q)$ l'ensemble des sous-espaces vectoriels de dimension q dans $F$ et

$$Q = \sum_{j=1}^{7} \widehat{\pi}_j \left\| \widehat{\theta}_j - \alpha_j - \beta - \gamma_j \right\|^2_{\widehat{\tau}}.$$

Si q est fixé, les estimateurs des moindres carrés asymptotiques de $\alpha_j$, $\beta$, $U$ et $\gamma_j$ sont donnés par

$$\left(\widehat{U}_q, \widehat{\beta}, \widehat{\alpha}_j, \widehat{\gamma}_j, j = \overline{1,7}\right) \in argmin\left\{Q, \alpha_j \in \mathbb{R},\ \beta \in F,\ \gamma_j \in U, U \in vect(F,q),\ q \in \mathbb{N}\right\}$$

sous les contraintes

$$\sum_{j=1}^{7} \widehat{\pi}_j \alpha_j = \widehat{\alpha}_0, \sum_{j=1}^{7} \widehat{\pi}_j \gamma_j = 0,$$

$$\int_0^1 \widehat{\tau}(x) \beta(x)\, dx = \widehat{\beta}_0, \int_0^1 \widehat{\tau}(x) \gamma_j(x)\, dx = 0.$$

On montre que :

**Proposition.** *Soit*

$$\begin{aligned}\widehat{\alpha}_j &= \int_0^1 \widehat{\tau}(x) \widehat{\theta}_j(x)\, dx - \widehat{\beta}_0 \\ \widehat{\beta} &= \sum_{j=1}^{7} \widehat{\pi}_j \widehat{\theta}_j - \widehat{\alpha}_0 \\ \widetilde{\theta} &= \widehat{\theta} - \widehat{\alpha}_j - \widehat{\beta}\end{aligned}$$

$$\begin{aligned} Q_1 &= \sum_{j=1}^{7} \widehat{\pi}_j \left\| \widehat{\theta}_j - \widehat{\alpha}_j - \widehat{\beta} - \gamma_j \right\|_{\widehat{\tau}}^2 \\ V &= \sum_{j=1}^{7} \widehat{\pi}_j \widetilde{\theta}_j \otimes \widetilde{\theta}_j \end{aligned}$$

*alors*

$$\begin{aligned} min\left\{Q, \alpha_j \in \mathbb{R},\ \beta \in F,\ \gamma_j \in U, U \in vect\left(F,q\right),\ q \in \mathbb{N}\right\} &= \\ min\left\{Q_1, \gamma_j \in U, U \in vect\left(F,q\right),\ q \in \mathbb{N}\right\} & \end{aligned}$$

*Soit* $\widehat{U}_q$ *le sous-espace vectoriel de* F *engendré par les vecteurs propres associés aux q valeurs propres les plus grandes de l'opérateur à noyau* $V\widehat{\tau}$. *Soit* $\widehat{\gamma}_j = \Pi_q \widetilde{\theta}_j$ *où* $\Pi_q$ *désigne le projecteur orthogonal sur* $\widehat{U}_q$. *Alors*

$$\left(\widehat{U}_q, \widehat{\gamma}_j, j = \overline{1,7}\right) \in argmin\left\{Q_1, \gamma_j \in U,\ \left(j = \overline{1,7}\right),\ U \in vect\left(F,q\right)\right\}$$

# Références


Baccini A.,Caussinus H. de Falguerolles A. (1993). Analysing dependence in large contingency tables : Dimensionality and patterns in scatter-plots. Multivariate analysis : future directions 2. Cuadras C.M. and Rao C. R. editors , pp. 245-263.

Dobson A. (1983). An introduction to Statistical Modellings. Chapman &Hall ed.

Dossou-Gbété S. (2001). Extension de la méthode des moindres carrés asymptotiques à l'estimation dans les modèles bilinéaires fonctionnels (en préparation).

Gourieroux Ch. et Monfort A. (1989). Statistique et modèles économétriques, vol. 1. Economica ed.

Green P.J. et Silverman B.W. (1994). Nonparametric regression and generalized linear models. A rougthness penalty approach. Chapman & Hall ed.

McCullagh P. et Nelder J.A. (1989). Generalized linear models. Chapman & Hall ed.

Ramsay J. O. & Silverman B.W. (1997). Functionnal Data Analysis. Springer Verlag ed.

Taylor W.F. (1953). Distance functions and regular best asymptotically Normal estimates, Annals of mathematical statistics, **24**, pp. 85-92.

# Quelques principes de déviations modérées et lois du logarithme itéré dans le modèle autorégressif hilbertien

**Ludovic MENNETEAU**

CREST et Université Montpellier 2
e-mail : mennet@stat.math.univ-montp2.fr

Exposé du 21 janvier 2002

## Résumé

Soit $\zeta = (\zeta_t)_{t\in\mathbb{R}}$, un processus à temps continu. Il est possible d'associer à $\zeta$ une suite infinie de processus,

$$X_n(t) = \zeta_{nT+t}, \text{ où } 0 \leq t \leq T \text{ et } n \in \mathbb{Z}.$$

$(X_n)_{n\in\mathbb{Z}}$ est alors une suite de variables aléatoires à valeurs fonctionnelles (par exemple dans $L^2([0,T])$, l'espace des fonctions de carré intégrable sur $[0,T]$). Pour faire de la prévision statistique sur les processus à temps continu, Bosq a considéré les processus linéaires fonctionnels (à valeurs dans un espace de Banach) parmi lesquels le modèle autorégressif hilbertien d'ordre 1, défini par l'équation

$$X_n = \rho(X_{n-1}) + \varepsilon_n,$$

où $(\varepsilon_k)_{k\in\mathbb{Z}}$ est une suite de variables aléatoires i.i.d. centrées à valeurs dans un espace de Hilbert $H$ et $\rho$ est un opérateur linéaire borné de $H$ dans $H$.
Dans cet exposé, je présente des résultats concernant la convergence de la covariance empirique

$$C_n = \frac{1}{n}\sum_{k=1}^{n} X_k \otimes X_k, \ (\text{ avec } x \otimes y : h \in H \mapsto < x, h > y)$$

associées à la suite $(X_n)$ vers l'opérateur de covariance de $X_0$

$$C = \mathbb{E}(X_0 \otimes X_0),$$

dans l'espace des opérateurs de Hilbert-Schmidt de $H$. Plus précisément, je présente un théoreme de déviations modérées et une loi du logarithme itéré pour $C_n - C$. Comme corollaire de ces résultats, j'obtiens des principes de déviations modérées et des lois du logarithme itérés relatifs aux éléments propres (valeurs propres et projecteurs associés) de $C$.

Une partie des résultats présentés est issue d'une collaboration avec André Mas (CREST et Université Toulouse 3).

# Références

H. Besse and H. Cardot, (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1, *Canad. Journal of Stat.* **24**, 467-487.
H. Besse, H. Cardot and D. Stephenson, (2000). Autoregressive forecasting of some climatic variations, *Scand. Journal of Stat.* **27** , 673-687.
D. Bosq, (2000). *Linear processes in function spaces*, Lecture Notes, Springer-Verlag.
J. Dauxois, A. Pousse and Y. Romain, (1982). Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *J. Multivar. Anal.* **12** , 136-154.
A. Mas, L. Menneteau, (2001). Large and moderate deviations for infinite dimensional autoregressive processes, (soumis).
L. Menneteau, (2001). Some laws of the iterated logarithm in hilbertian autoregressive models, (soumis).

# Nonparametric estimation applied to sismicity of Galicia

**Alejandro QUINTELA DEL RIO ***
En collaboration avec Graciella ESTÉVEZ PÉREZ

* Adresse pour correspondance :
Departamento de Matemàticas, Facultad de Informatica
Universidad de A Coruña, Campus de Elviña
A Coruña, Espagne
e-mail : eiuqinte@udc.es

Premier exposé du 4 Février 2002

---

## Abstract

We make use of nonparametric estimation of hazard and intensity functions to describe the temporal structure of seismic activity. Kernel estimation of hazard function has confirmed that earthquakes tend to grouping and this statistical application has been used to study the occurrence process of the clusters formed. Kernel intensity estimation has helped us to describe the occurrence process of cluster members. We compare two geographic areas of Spain (Granada and Galicia), the first more studied by geological research along the years, and we can conclude that the seismic activity in these two regions is not very different.

# A modification of cross-validation procedure in kernel hazard estimation from dependent samples

**Graciela ESTÉVEZ PÉREZ ***
En collaboration avec Alejandro QUINTELA DEL RIO et Philippe VIEU

* Adresse pour correspondance :
Departamento de Matemàticas, Facultad de Ciencias
Campus de A Zapateira 76, Universidad de A Coruña
15071 A Coruña, Espagne
e-mail : graci@udc.es

Second exposé du 4 Février 2002

## Abstract

This talk is devoted to the nonparametric estimation of hazard function by means of kernel smoothers. Rates of convergence for kernel hazard smoothers have been studied a lot in the literature (see for instance Estevez, 2001, for recent results).

In this talk we will concentrate more specifically on the crucial problem of bandwidth selection. We first get the convergence rate of usual cross-validated bandwidth under a general dependence assumption on the sample data (Hart and Vieu, 1990 ; Youndjé et al., 1996 ; Estévez and Quintela, 1999), extending in several directions the results existing in the literature (Hall and Marron, 1987). In a second attempt, this rate of convergence is used to motivate the introduction of a penalized version of the cross-validation procedure, as suggested in Estévez et al., (2001). The rate of convergence is calculated, and a short simulation study shows the interest of this approach for finite sample studies. Finally, as a by - product of our proofs, we state a general inequality for the moments of sums of strong dependent variables, which is an extension of a similar result given in Kim and Cox (1995).

The results obtained for this bandwidth selection problem in hazard estimation will be of great interest for practical applications such as the sismologic studies discussed in the previous talk (see Estévez and Quintela, 2002).

# Références

Estévez, G. (2001). On convergence rates for quadratic errors in kernel hazard estimation. *Preprint.*

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

Estévez, G. and Quintela, A. (2002). Estimación no paramétrica de la función de riesgo : aplicaciones a sismología. *Questiió.* vol. 26 (in press).

Estévez, G., Quintela, A. and Vieu, P. (2001). Convergence rate for cross-validatory bandwidth in kernel hazard estimation from dependent samples. *J. Statist. Plan. Inference* (in press).

Hall, P. and Marron, J.S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Relat. Fields,*, **74**, 567-581.

Hart, J. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, **18**, 873-890.

Kim, T.Y. and Cox, D.D. (1995). Asymptotic behaviors of some measures of accuracy in nonparametric curve estimation with dependent observations. *J. Multivariate Anal.*, **53**, 67-93.

Youndjé, É. ; Sarda, P. and Vieu, P. (1996). Optimal smooth hazard estimates. *TEST*, **5**, 379-394.

# Partially linear models with dependent errors : some notes on estimation, bandwidth selection and testing of hypotheses

**German ANEIROS PÉREZ**

Departamento de Matemàticas, Facultad de Informatica
Universidad de A Coruña, Campus de Elviña
A Coruña, Espagne
e-mail : ganeiros@udc.es

Exposé du 11 Février 2002

## Abstract

Let us consider the partially linear regression model

$$y_i = \zeta_i^T \beta + m(t_i) + \varepsilon_i,\ i = 1, ..., n,$$

where the $(p \times 1)$ vector $\beta$ and the function $m$ are unknown, where $\zeta_i$ and $t_i$ are design points (random and fixed, respectively) and where the random errors $\{\varepsilon_i\}$ are stationary and strong mixing. This model is interesting in many fields. In the field of economics, generally the focus of attention is the estimation of $\beta$, being more important than the behavior of $m$.

In this talk, we will firstly discuss the asymptotic normality of the Least Squares estimator of $\beta$ (based on kernel type estimation) under strong mixing errors. This estimate was proposed and studied by Robinson (1988) and Speckman (1988) in the setting of independent errors. under strong mixing errors recent results were given in Aneiros-Perez (2001b), while Aneiros-Perez and Quintela del Rio (2001b) give similar results for a Generalized Least Squares estimator of $\beta$.

Then, by means of second order approximations for the variance and bias, we obtain the expression of an asymptotically optimal bandwidth for the LS estimator (thus, we generalize the result of Linton (1995), who worked under independent errors). Recent advances in bandwidth selection under this partially

linear model can be found in Aneiros-Perez (2000), (2001a) and (2001b), and in Aneiros-Perez and Quintela del Rio (2001a) and (2002).

Then, we will consider the problem of testing

$$H_0 : \beta = \beta_0$$

The asymptotic distributions of the corresponding test statistic are obtained under $H_0$ and also under local alternatives and under fixed alternatives (thus, we generalize the result of Gao (1997), who worked under independent errors). These results can be found in Gonzalez-Manteiga and Aneiros-Perez (2001).

Finally results for the case of estimation of $m$ will be given, conditioned on $\zeta_i^T$. A short simulation study will be carried out.

# Références

Aneiros Pérez, G. (2000). Bandwidth selection in kernel smoothing of the nonparametric part in a partial linear model with autorregressive errors. *Qüestiió*, **24**, 267-291, (in Spanish).
Aneiros Pérez, G. (2001a). Plug-In Bandwidth Choice for Estimation of Nonparametric Part in Partial Linear Regression Models with Strong Mixing Errors *Preprint*.
Aneiros Pérez, G. (2001b). On bandwidth selection in partial linear regression models under dependence *Preprint*.
Aneiros Pérez, G. and Quintela del Río, A. (2001a). Modified Cross-Validation in Semiparametric Regression Models with Dependent Errors. *Communications in Statistics : Theory and Methods*, **30**, 289-307.
Aneiros Pérez, G. and Quintela del Río, A. (2001b). Asymptotic properties in partial linear models under dependence. To appear in *Test*, **10**, No. 2.
Aneiros Pérez, G. and Quintela del Río, A. (2002). Plug-in bandwidth choice in partial linear models with autoregressive errors. *Journal of Statistical Planning and Inference*, **100**, 23-48.
Gao, J. (1997). Adaptive parametric test in a semiparametric regression model, *Commun. Statist.-Theory Meth.* **26**, 787-800.
González Manteiga, W. and Aneiros Pérez, G. (2001). Testing in partial linear regression models with dependent errors. *Preprint* .
Linton, O. (1995). Second order approximation in the partially linear regression model, *Econometrica* **63**, 1079-1112.
Robinson, P. (1988). Root-N-consistent semiparametric regression, *Econometrica* **56**, 931-954.
Speckman, P. (1988). Kernel smoothing in partial linear models, *J. Roy. Statist. Soc., Ser. B*, **50**, 413-436.

# Sélection des variables en régression linéaire ; lien avec le modèle linéaire fonctionnel

**Guy Martial NKIET**

Université de Masuk
Gabon
e-mail : gnkiet@hotmail.com

Second exposé du 25 Février 2002

## Résumé

Nous considérons le modèle de régression $Y = \sum_{i=1}^{p} \alpha_i X_i + \varepsilon$ où les $X_i$ sont des v.a. réelles centrées et linéairement indépendantes, et $\varepsilon$ est une v.a.r centrée indépendante de $X = (X_1, \cdots, X_p)$. Nous nous intéressons au problème de sélection des variables dans ce modèle, c'est-à-dire à l'estimation de l'ensemble $I_1 = \{1 \leq i \leq,\ \alpha_i \neq 0\}$ supposé non vide. Pour cela, nous utilisons un critère d'invariance de l'analyse canonique linéaire pour les transformations $(Y, X) \mapsto (Y, A_K X)$, où $K \subset I := \{1, \cdots, p\}$ et $A_K : x \in \mathbb{R}^p \mapsto (x_i)_{i \in K} \in \mathbb{R}^{card(K)}$. Ce critère, obtenu à partir des résultats de Dauxois et Nkiet (1997), est défini par $C_K = \left\| V_{12} - V_1 A_K^* \left(A_K V_1 A_K^*\right)^{-1} A_K V_{12} \right\|$ où $V_{12} := \mathbb{E}(YX)$, $V_1 := \mathbb{E}(X \otimes X)$ et $\|\cdot\|$ est une norme de $\mathbb{R}^p$. On montre d'abord que l'on a $C_K = 0$ si, et seulement si $I_1 \subset K$, ce qui indique que l'on doit minimiser ce critère. Celui-ci n'étant pas connu en pratique, on se base sur un estimateur convergent. Lorsque l'on a un échatillon iid $\left(Y_k, X^{(k)}\right)_{1 \leq k \leq n}$ de $(Y, X)$, on considère $V_{12}^{(n)} := n^{-1} \sum_{k=1}^{n} Y_k X_k$, $V_1^{(n)} := n^{-1} \sum_{k=1}^{n} X_k \otimes X_k$, et on estime $C_K$ par $C_K^{(n)} := \left\| V_{12}^{(n)} - V_1^{(n)} A_K^* \left(A_K V_1^{(n)} A_K^*\right)^{-1} A_K V_{12}^{(n)} \right\|$.

Pour estimer $I_1$, on cherche d'abord une caractérisation de cet ensemble ; on considère $K_i := I - \{i\}$ et la permutation $\sigma$ de $I$ telle que $C_{K_{\sigma(1)}} \geq C_{K_{\sigma(2)}} \geq \cdots \geq C_{K_{\sigma(p)}}$, les ex-aequo étant ordonnés dans le sens croissant des indices correspondant ; on a alors $I_1 = \{\sigma(i),\ 1 \leq i \leq q_0\}$, où $q_0$ est l'unique élément de $I$ vérifiant $C_{K_{\sigma(q_0)}} > 0$ et $C_{K_{\sigma(q_0+1)}} = 0$. L'estimation de $I_1$ se ramène alors à celles de $\sigma$ et $q_0$. Pour estimer $\sigma$, on introduit une suite $(k_n)_{n \geq 1}$ de fonctions de $I$ vers $\mathbb{R}$ pour laquelle il existe une fonction strictement décroissante $k$ et un réel $\beta \in ]0, 1/2[$ telles que $\lim_{n \to +\infty} n^{\beta} k_n(l) = k(l)$ ; on définit alors $\psi_l^{(n)} := C_{K_l}^{(n)} + k_n(l)$ et on considère

la permutation aléatoire $\sigma^{(n)}$ de $I$ telle que $\psi^{(n)}_{\sigma^{(n)}(1)} \geq \psi^{(n)}_{\sigma^{(n)}(2)} \geq \cdots \geq \psi^{(n)}_{\sigma^{(n)}(p)}$, les ex-aequo étant ordonnés dans le sens croissant des indices correspondant ; on montre alors que l'on a $\lim_{n\to+\infty} P\left(\sigma^{(n)} = \sigma\right) = 1$. Pour estimer $q_0$, on pose $J^{(n)}_l := \left\{\sigma^{(n)}(1), \cdots, \sigma^{(n)}(l)\right\}$, on introduit une suite $(h_n)_{n\geq 1}$ de fonctions de $I$ vers $\mathbb{R}$ pour laquelle il existe une fonction strictement croissante $h$ et un réel $\alpha \in ]0, 1/2[$ telles que $\lim_{n\to+\infty} n^\alpha h_n(l) = h(l)$, et on définit $\varphi^{(n)}_l := C^{(n)}_{J^{(n)}_l} + h_n(l)$. On estime alors $q_0$ par $q^{(n)}_0 = \arg\min_{l\in I}\left(\varphi^{(n)}_l\right)$ et on prouve que l'on a en probabilité $q_0 = \lim_{n\to+\infty} q^{(n)}_0$. La sélection des variables est effectuée en prenant $I^{(n)}_1 = \left\{\sigma^{(n)}(i),\ 1 \leq i \leq q^{(n)}_0\right\}$ ; les résultats de convergence obtenus impliquent $\lim_{n\to+\infty} P\left(I^{(n)}_1 = I_1\right) = 1$, ce qui signifie que la méthode proposée est convergente.

Les outils utilisés étant encore définis lorsque l'on considère des v.a. hilbertiennes, une extension de la méthode précédente à la sélection des variables dans le modèle linéaire fonctionnel est envisageable.

# Références

Dauxois, J., Nkiet, G.M. (1997). Canonical analysis of two Euclidean subspaces and its applications, *Linear Algebra Appl.*, **264**, 355-388.

Nkiet G.M., (2002). Inference for the invariance of canaonical analysis under linear transformations, *J. Multivariate Anal.*, to appear.

Thompson M.L. (1978). Selection of variables in multiple regression, Part I, *Internat. Statist. Rev.*, **46**, 1-19.

Zheng X., Loh W.Y. (1997). A consistent variable selection criterion for linear models with high-dimensional covariates, *Statist. Sinica*, **7**, 311-325.

# ACP Banachique

**Tawfik BENCHIKH**
En collaboration avec A. YOUSFATE

Université de Sidi Bel Abbés

Exposé du 3 Avril 2002

**Keywords** A.C.P., dualité, espace de Banach, opérateur compact, ordre spectral, projection, base de Schauder

## Résumé

Pour appliquer l'*A.C.P.* dans des espace de dimension infinie ( traitement du signal, traitement de l'image, fonction aléatoires, ...), le vecteur associé à un individu devient une fonction ( ou une trajectoire ). Si les fonctions étudiées sont dans $L^2$, les travaux de Dauxois et Pousse donnent la solution de l'*A.C.P.* sous forme globale ou sous forme itérative. Dans ce travail nous généralisons l'*A.C.P.* à des fonctions dans un espace de Banach réel séparable et réflexif sans faire subir une transformation aux fonctions de base pour obtenir des solutions linéaires de l'*A.C.P.*.

La définition de l'*A.C.P.* banachique est donnée par :
on appelle *A.C.P.* "pas-à-pas" de $U$, où $U$ est un opérateur d'un espace de Banach $E$ à valeur dans un espace de Banach F, le processus itératif d'optimisation suivant (ainsi que les conséquences des résultats qui en découlent) :

$$\begin{cases} max_{e\in E}\langle \Phi e, U\circ\Psi\circ U'\circ\Phi e\rangle \\ \|e\| = 1; \end{cases} \tag{1}$$

Si $e_1$ est un argument de solution, on note $e_1^{'} = \Phi(e_1)$, le deuxième argument de solution doit vérifier la solution du problème suivant :

$$\begin{cases} max_{e\in E}\langle \Phi e, U\circ\Psi\circ U'\circ\Phi e\rangle \\ \|e\| = 1 \quad \text{et} \quad \langle e_1^{'}, e\rangle = 0. \end{cases} \tag{2}$$

et itération sous contrainte de $*$-orthogonalité des arguments des solutions $(e_i)_{i\in I}$ aux $[\Phi e_1, \ldots, \Phi e_{i-1}]$. $\Phi$ (resp. $\Psi$) est un opérateur linéaire de $E$ dans son dual topologique $E'$ (resp. De $F$ dans $F'$) et $U'$ le transposé de $U$

# Références

Benchikh, T. (1999). Analyses factorielles dans un espace de Banach sous contraintes linéaires", Magister, Sidi-Bel-Abbès.
Brezis, H. (1987). *Analyse Fonctionnelle, Théorie et Applicati ons.* 2e Edition, MASSON, Paris.
Dauxois, J. et Pousse, A. (1976). *'Les analyses factorielles e n calcul des probabilité et en statistique : essai d'etude synthétique.* Thèse es-sciences , Toulouse.
Kato T. (1966). *Pertubation Theory for Linear Operators.* Spr ing-Verlag.

# ACP dans le domaine des fréquences

**Alain BOUDOU et Sylvie VIGUIER-PLA**

Laboratoire de Statistique et Probabilités
Université Paul Sabatier
31062 Toulouse, France
e-mail : boudou@cict.fr et viguier@cict.fr

Exposé du 25 Mars 2002

## 1. Introduction

L'analyse dans le domaine des fréquences d'une série stationnaire $p$ dimensionnelle (c'est-à-dire d'une famille $(X_n)_{n\in\mathbb{Z}}$ d'éléments de $L^2_{\mathbb{C}^p}(\Omega,\mathcal{A},\mu)$ telle que $\mathbb{E}X_n^t\overline{X_m} = \mathbb{E}X_{n-m}^t\overline{X_0}$) développée en B.D. et Br. nécessite l'analyse en composantes principales (A.C.P.) de chacune des composantes spectrales. Se ramenant donc à la diagonalisation d'une infinité de matrices elle ne peut donc être envisagée d'une façon concrète. Ici nous proposons, grâce à une discrétisation du spectre, de tourner cette difficulté. Nous étudions une condition suffisante (liée à la fonction de densité spectrale) de convergence de la solution approchée ainsi obtenue. Moyennant une hypothèse d'ergodicité nous donnons une estimation des opérateurs qu'il est alors nécessaire de connaître. Enfin, nous examinons un exemple concernant les températures moyennes mensuelles de plusieurs villes de France.

## 2. Notations

Lorsque $Z$ est une mesure aléatoire $p-$dimensionnelle ($p-$m.a.), c'est-à-dire une mesure vectorielle définie sur la tribu $\mathcal{B}$ des boréliens de $[-\pi,\,\pi[$ à valeurs dans $L^2_{\mathbb{C}^p}(\Omega,\mathcal{A},\mu)$ telle que $\mathbb{E}\,ZA^t\overline{ZB} = 0$ lorsque $A$ et $B$ sont disjoints, une relation d'équivalence, liée à la mesure $M_Z : A\in\mathcal{B}\mapsto \mathbb{E}\,ZA^t\overline{ZB}\in\mathcal{L}(\mathbb{C}^p)$, est définie sur un sous-espace vectoriel de $\mathcal{L}(\mathbb{C}^p,\mathbb{C}^q)^{[-\pi,\,\pi[}$.

L'ensemble des classes d'équivalence ainsi obtenues, noté $(p,q)-L^2(M_Z)$ et appelé "ensemble des fonctions de carré $M_Z-$intégrable, possède une structure

d'espace de Hilbert. L'intégrale stochastique, relativement à $Z$, peut se définir comme une isométrie de $(p,q)-L^2(M_Z)$ sur $\overline{\text{vect}}\{K\circ ZA;\ A\in\mathcal{B}, K\in\mathcal{L}(\mathbb{C}^p,\mathbb{C}^q)\}$.

A toute suite stationnaire $p-$dimensionnelle $(X_n)_{n\in\mathbb{Z}}$ on peut associer une et une seule mesure aléatoire $p-$dimensionnelle $Z^X$ telle que $X_n=\int \mathrm{e}^{i.n}I\,\mathrm{d}Z^X$.

L'image de $(X_n)_{n\in\mathbb{Z}}$ par le filtre $\varphi$, élement de $(p,q)-L^2(M_Z)$, est la série $q-$dimensionnelle stationnaire $(\int \mathrm{e}^{i.n}\varphi(.)\,\mathrm{d}Z^X(.))_{n\in\mathbb{Z}}$ dont la mesure aléatoire $q-$dimensionnelle associée est $Z_q^X\ : A\in\mathcal{B}\mapsto\int \mathbb{1}_A(.)\varphi(.)\mathrm{d}Z^x(.)\in L^2_{\mathbb{C}^p}(\Omega,\mathcal{A},P)$.

# 3. L'Analyse

Le but de l'analyse est de résumer une série stationnaire $p-$dimensionnelle $(X_n)_{n\in\mathbb{Z}}$ par une série stationnaire $q-$dimensionnelle $(Y_n)_{n\in\mathbb{Z}}$ stationnairement corrélée avec $(X_n)_{n\in\mathbb{Z}}$. Afin dévaluer les qualités du résumé $(Y_n)_{n\in\mathbb{Z}}$, on transforme cette dernière, par filtrage, en une série $p-$dimensionnelle $(\int \mathrm{e}^{i.n}\varphi(.)\,\mathrm{d}Z^Y)_{n\in\mathbb{Z}}$. La quantité $\|X_n-\int \mathrm{e}^{i.n}\varphi(.)\,\mathrm{d}Z^Y(.)\|$ est indépendante de $n$ du fait des propriétés de stationnarité. C'est la raison pour laquelle $\inf\{\|X_0-\int\varphi(.)\,\mathrm{d}Z^Y(.)\|;\varphi\in(q,p)-L^2(M_{Z^Y})\}$ mesure la qualité du résumé $(Y_n)_{n\in\mathbb{Z}}$. D'où la

**Définition**. L'A.C.P. d'ordre $q$ de $(X_n)_{n\in\mathbb{Z}}$ est la recherche d'une série stationnaire $q-$dimensionnelle $(Y_n)_{n\in\mathbb{Z}}$, stationnairement corrélée avec $(X_n)_{n\in\mathbb{Z}}$ et d'un élément $\varphi$ de $(q,p)-L^2(M_{Z^Y})$ de telle sorte que $\|X_0-\int\varphi\mathrm{d}\,Z^Y\|$ soit le plus petit possible.

C'est-à dire que parmi toutes les séries possibles nous choisissons le "meilleur résumé".

Notant $Z^X$ la mesure aléatoire $p-$dimensionnelle associée à $(X_n)_{n\in\mathbb{Z}}$, si $\mu$ est une mesure dominant $t_{Z^X}=\mathrm{tr}\,M_{Z^X}(.)$ et si $\sum_{j=1}^n\lambda_j(.)a_j(.)\otimes a_j(.)$ est une décomposition de Schmidt "mesurable" de $\frac{\mathrm{d}M_{Z^X}}{\mathrm{d}\,\mu}$, la série optimale est $(\int \mathrm{e}^{i.n}\sum_{j=1}^q a_j\otimes f_j\mathrm{d}\,Z^X)_{n\in\mathbb{Z}}$ où $\{f_1,\ldots,f_q\}$ est la base canonique de $\mathbb{C}^q$.

Cela nécessite la diagonalisation de $\frac{\mathrm{d}M_{Z^X}}{\mathrm{d}\,\mu}(\lambda)$ pour tout $\lambda$ de $[-\pi,\pi[$ et ne peut donc être envisagé d'un point de vue pratique.

# 4. Discrétisation du spectre

$k$ étant un élement quelconque de $\mathbb{N}^*$, posons :

$A_{kk}=\{-\pi\}$, $A_{0k}=]-\frac{\pi}{k},\frac{\pi}{k}[$, $A_{nk}=]\frac{\pi\,n}{k}-\frac{\pi}{k},\frac{\pi\,n}{k}]$ pour $n=-k+1,\ldots,-1$ et $A_{nk}=[\frac{\pi\,n}{k},\frac{\pi\,n}{k}+\frac{\pi}{k}[$ pour $n=1,2,\ldots,k-1$.

$\{A_{nk},n=-k+1,\ldots,k-1\}$ constituant une partition de $[-\pi,\pi[$, à l'infinité d'analyses spectrales (A.S.) nécessaire à la mise en oeuvre de l'A.C.P. de $Z^X$ (ou de $(X_n)_{n\in\mathbb{Z}}$), il paraît naturel de substituer l'A.S. de chacune des $2\,k$ matrices $M_{Z^X}(A_{nk})$. Il s'agit en fait de l'A.C.P. de $\mathcal{L}_k(Z^X)$, $p-$m.a. image de $Z^X$ par

l'application $\mathcal{L}_k = \sum_{n=-k+1}^{k-1} \frac{\pi}{k} \mathbb{1}_{A_{nk}}$.

Notons $\alpha_k$ l'élément de $(p,q)-L^2(M_{\mathcal{L}_k(Z^X)})$ correspondant à l'A.C.P. d'ordre $q$ de la $p-$m.a. $\mathcal{L}_k(Z^X)$ (qu'il est donc possible d'obtenir grâce à la diagonalisation de $2\,k$ matrices).

Si l'on fait l'hypothèse que $\alpha$, *application de* $[-\pi,\,\pi[$ *dans* $HS(p,q)$, *est continue*, on peut démontrer que :

$$\begin{aligned}&\lim_k\inf\{\|X_0 - \int \varphi \mathrm{d}(Z^X)_{\alpha_k(\mathcal{L}_k)}\|; \varphi \in (q,p) - L^2(M_{(Z^X)_{\alpha_k(\mathcal{K}_k)}}\}\\ &= \inf\{\|X_0 - \int \varphi \mathrm{d}(Z_\alpha)\|; \varphi \in (q,p) - L^2(M_{Z_\alpha})\}\ .\end{aligned}$$

Egalité qui nous permet de voir que la qualité du résumé qu'est $(\int \mathrm{e}^{i.n}\alpha_k(\mathcal{L}_k)\mathrm{d}Z^X)_{n\in\mathbb{Z}}$ de $(X_n)_{n\in\mathbb{Z}}$ peut être aussi proche qu'on le souhaite, pourvu que l'on choisisse $k$ suffisamment grand, de la qualité du résumé optimal $(\int \mathrm{e}^{i.n}\alpha \mathrm{d}Z^X)_{n\in\mathbb{Z}}$. L'hypothèse effectuée légitime en quelque sorte la discrétisation du spectre. Elle est vérifiée lorsque la *fonction d'autocovariance est absolument sommable (c'est-à-dire lorsque* $\sum_{n\in\mathbb{Z}} \|\mathbb{E} X_n^t \overline{X_0}\| < +\infty$*) et lorsque, pour tout* $\lambda$ *de* $[-\pi,\,\pi[$, $F(\lambda)$, *valeur de la densité spectrale en* $\lambda$, *possède* $p$ *valeurs propres distinctes.*

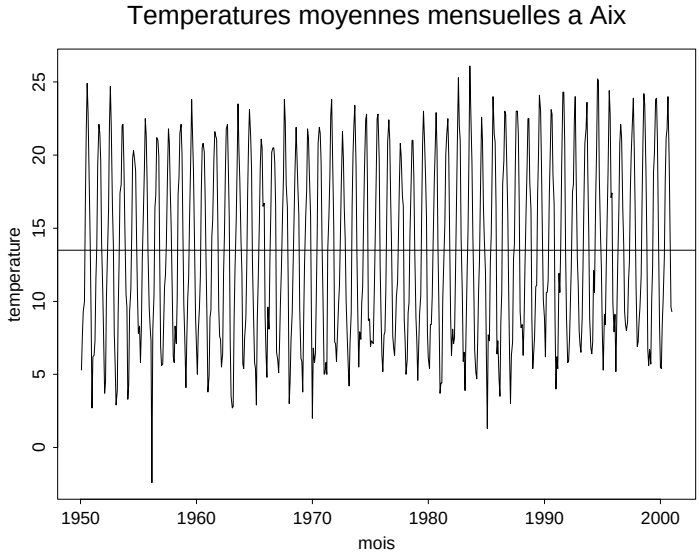
Si de plus on fait l'hypothèse que $(X_n)_{n\in\mathbb{Z}}$ *est stationnaire au sens strict et ergodique*, désignant par $I_m(\lambda)$ la matrice aléatoire $(2\,\pi\,m)^{-1}(\sum_{l=1}^m \mathrm{e}^{-i\,\lambda\,l}X_l)^t\overline{(\sum_{l=1}^m \mathrm{e}^{-i\,\lambda\,l}X_l)}$, on peut démontrer que $\int_{A_{nk}}(I_m(\lambda))(\omega)\mathrm{d}\,\eta(\lambda)$ est un estimateur de $M_{Z^X}(A_{nk})$ convergeant presque sûrement.

# 5. Les données d'application

On se propose d'appliquer la méthode d'A.C.P. exposée plus haut aux températures moyennes mensuelles de 16 villes de France, de jan1950 à dec2000

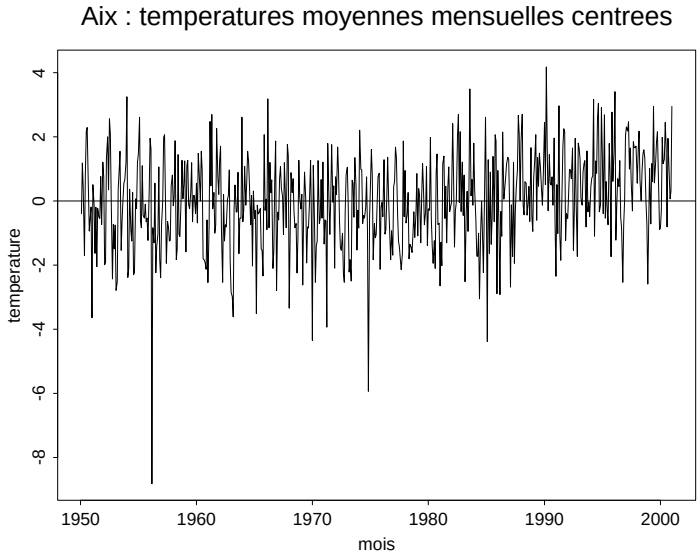
| | aix | biarritz | blagnac | ... | paris14e | perpignan | reims | rennes | strasbourg |
|---|---|---|---|---|---|---|---|---|---|
| jan1950 | 5.3 | 6.9 | 4.3 | ... | 3.2 | 7.4 | 1.0 | 4.2 | -0.1 |
| fev1950 | 7.6 | 11.4 | 8.4 | ... | 7.9 | 9.9 | 6.1 | 7.8 | 5.3 |
| mar1950 | 9.3 | 9.5 | 8.1 | ... | 8.9 | 11.4 | 6.3 | 8.8 | 7.2 |
| avr1950 | 10.0 | 11.1 | 10.1 | ... | 9.8 | 12.5 | 8.2 | 9.4 | 9.0 |
| mai1950 | 16.7 | 15.3 | 15.7 | ... | 15.4 | 17.2 | 13.4 | 14.0 | 15.6 |
| ... | | | | | | | | | |
| sep2000 | 19.7 | 19.4 | 20.0 | ... | 17.5 | 21.3 | 16.5 | 17.5 | 16.7 |
| oct2000 | 14.7 | 14.9 | 14.1 | ... | 12.6 | 16.5 | 12.0 | 12.4 | 12.4 |
| nov2000 | 9.6 | 11.9 | 10.2 | ... | 8.9 | 11.9 | 8.0 | 9.4 | 7.4 |
| dec2000 | 9.3 | 12.8 | 9.9 | ... | 7.8 | 10.9 | 6.6 | 8.8 | 5.6 |

Pour une de ces villes, voici ces températures moyennes mensuelles :




# 6. Centrage avant ACP

Pour obtenir des données quasi-stationnaires, on “centre” les données par rapport à la moyenne des moyennes mensuelles, ce qui donne, pour la même ville que prédemment :




# 7. A.C.P.

i). La méthode.

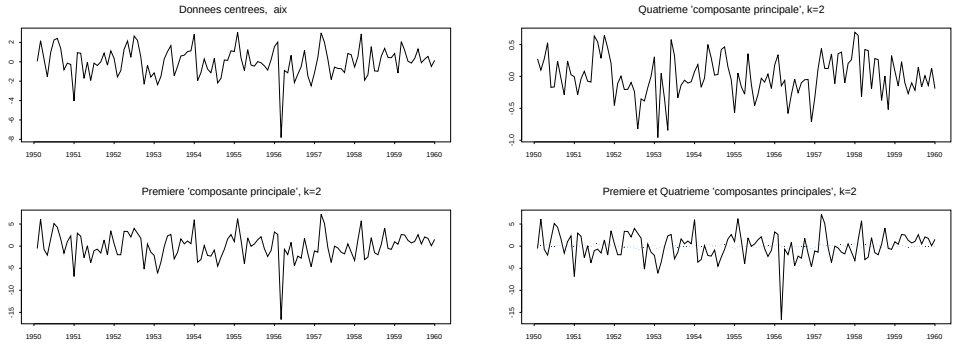
Rappelons que nous approchons $M_{ZX}(A_{nk})$ par la matrice aléatoire
$(2\,\pi\,m)^{-1}(\sum_{l=1}^{m} \mathrm{e}^{-i\,\lambda\,l} X_l)^t(\overline{\sum_{l=1}^{m} \mathrm{e}^{-i\,\lambda\,l} X_l})$,
dont une réalisation, basée sur les observations de $(X_n)$ rangées en lignes dans la matrice $X_{i=1,\ldots,m,j=1,\ldots,p}$, est notée
$T_{m,n,k} = \frac{1}{2\,\pi\,(m-1)} \sum_{l=1}^{m-1} \sum_{j=1}^{m-1} epi(n,k,l-j) X_l\, {}^t X_j$ où $epi(n,k,l) = \int_{A_{nk}} e^{i\lambda\,l} \mathrm{d}\lambda$.
Rappelons aussi
$A_{nk} = ]\frac{\pi}{k}(n-1),\ \frac{\pi}{k}n]$ si $n = -k+1, \ldots, -1$, $A_{0k} = ]-\frac{\pi}{k}, \frac{\pi}{k}[$ et $A_{nk} = [\frac{\pi}{k}n,\ \frac{\pi}{k}(n+1)[$ si $n = 1, \ldots, k-1$,
Pour simplifier les calculs, on peut utiliser la propriété : $T_{m,n,k} = \overline{T_{m,-n,k}}$

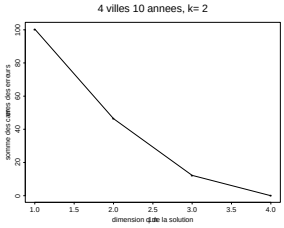
La méthode d'A.C.P. consiste donc en
– la recherche des vecteurs propres $a_{jnk}$ de $T_{m,n,k}$ : $T_{m,n,k} = \sum_{j=1}^{v} \lambda_{jnk} a_{jnk}\, {}^t\overline{a_{jnk}}$

– le calcul des coefficients de la "combinaison linéaire" qui permettra de calculer $X$ "projeté" sur les "axes principaux", c'est-à-dire $X$ réduit :
$C'_{lk} = \frac{1}{2\,\pi} \sum_{n=-k+1}^{k-1} epi(n,k,l) \sum_{j=1}^{q} a_{jnk} \otimes f_j$ où $f_j$ est la base canonique de $C^q$.

ii). Les résultats.

Les graphiques des résumés suivants montrent que le résumé d'ordre 1 est fidèle à l'allure des séries de départ. La variabilité des résumés est décroissante avec l'ordre.



L'examen des valeurs des $C'_{lk}$ motre une décroissance de ces coefficients quand $l$ grandit en valeur absolue :

pour $k = 2$, $C'_{0k}$=

```
          [,1]       [,2]        [,3]        [,4]
[1,] 0.4571209 -0.6370731 -0.08030861  0.05133392
[2,] 0.4814932  0.6444050 -0.05786562  0.03379486
[3,] 0.5237155  0.2794855  0.11433632  0.04233850
[4,] 0.5331499 -0.3077182 -0.02181226 -0.12752352
```

(lecture : la première colonne contient les coefficients de $X_i$ pour résumer $X_i$ sur le premier axe principal de l'A.C.P., la colonne $j$ pour le $j^{ème}$ axe principal).

$C'_{-119k}$ =

```
              [,1]          [,2]          [,3]          [,4]
[1,] -5.919915e-05  7.512154e-05 -0.0030217677 -0.0014125100
[2,]  5.690072e-05  1.203660e-04 -0.0002560523 -0.0028681844
[3,]  2.512979e-05 -2.718778e-05 -0.0014240541  0.0036967779
[4,]  5.046111e-05 -2.647828e-04  0.0041476740  0.0001404577
```

Lecture : la première colonne contient les coefficients de $X_{i-119}$ pour résumer $X_i$ sur le premier axe principal de l'A.C.P., la colonne $j$ pour le $j^{ème}$ axe principal).

Comparaison des sommes des carrés des erreurs quand $k$ varie :

| $q$ | $k=2$ | $k=4$ | $k=6$ | $k=7$ | $k=10$ |
|---|---|---|---|---|---|
| 1 | 100.420 | 97.642 | 94.281 | 87.800 | 78.568 |
| 2 | 46.390 | 41.606 | 31.226 | 30.789 | 28.193 |
| 3 | 12.249 | 10.166 | 7.910 | 8.053 | 6.801 |
| 4 | 0 | 0 | 0 | 0 | 0 |

$k = 2$, 10 années, 4 villes : corrélations série de départ centrée avec "composantes principales" :

```
            aix biarritz   blagnac  bourges    xnr1q4k2    xnr2q4k2    xnr3q4k2    xnr4q4k2
     aix  1.00000  0.82462  0.891951  0.88810  9.372e-001 -2.832e-001  0.01311674  0.01664755
biarritz  0.82462  1.00000  0.946237  0.87902  9.527e-001  2.768e-001 -0.01621212  0.01929959
 blagnac  0.89195  0.94624  1.000000  0.90472  9.777e-001  1.142e-001  0.05555424 -0.00666491
 bourges  0.88810  0.87902  0.904716  1.00000  9.618e-001 -1.196e-001 -0.05891242 -0.02444414
xnr1q4k2  0.93717  0.95270  0.977697  0.96175  1.000e+000 -2.259e-006 -0.00004513  0.00007256
xnr2q4k2 -0.28320  0.27684  0.114157 -0.11958 -2.258e-006  1.000e+000  0.00078625  0.00025200
xnr3q4k2  0.01312 -0.01621  0.055556 -0.05891 -4.513e-005  7.863e-004  1.00000000 -0.00028894
xnr4q4k2  0.01665  0.01930 -0.006677 -0.02444  7.256e-005  2.520e-004 -0.00028894  1.00000000
```

`xnriq4k2` est la notation pour la composante principale `i`.

Les corrélations sont plus fortes avec la première composante principale, qui restitue bien une grosse partie de l'ifnformation. On note que la ville d'Aix est la moins bien reconstituée des quatre.

# 8. Comparaison avec l'A.C.P. dans le cas où il y a périodicité

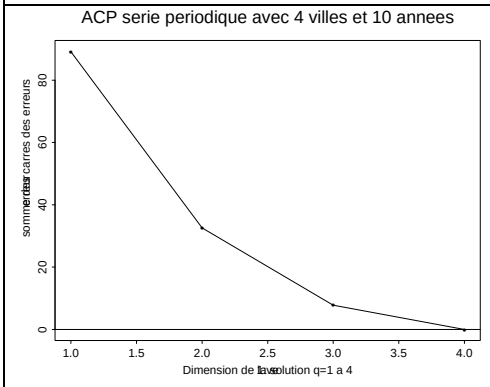
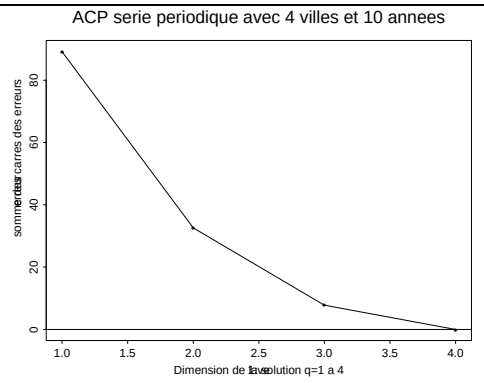
$Z$ = transformée de Fourrier de $X$, considérée comme les 51 mesures d'une v.a. de $\mathbb{R}^{16}$ périodique de période 12.

Procédure : ACP des 12 composantes spectrales $Z_i$, puis reconstitution dans le domaine temporel de $\widehat{X}$ par transformée de Fourrier inverse.

$Z_j = \sum_{k=1}^{12} \mathrm{e}^{-i\,k\,\pi/6} X_k$ où $X_k$ contient les $k-$ième mois de chaque année.

ACP de chaque $Z_j$, puis reconstitution $\widehat{X}_j = \sum_{k=1}^{12} \mathrm{e}^{i\,k\,\pi/6}\,\widehat{Z}_k$

somme des carrés des erreurs $= \sum_{j,k,l} (X_{j,k,l} - \widehat{X}_{j,k,l})^2$.

| Résultat pour les 51 années : | Résultat partiel : |
|---|---|
|  |  |



Comparaison des sommes des carrés des erreurs avec méthode précédente :

| $q$ | $k=2$ | $k=4$ | $k=6$ | $k=7$ | $k=10$ | méthode avec périodicité |
|---|---|---|---|---|---|---|
| 1 | 100.420 | 97.642 | 94.281 | 87.800 | 78.568 | 89.145 |
| 2 | 46.390 | 41.606 | 31.226 | 30.789 | 28.193 | 32.580 |
| 3 | 12.249 | 10.166 | 7.910 | 8.053 | 6.801 | 7.770 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |

La méthode exposée fait mieux que la méthode avec hypothèse de périodicité dès que $k$ dépasse 6. Cela peut s'expliquer par le fait que l'on crée $2k-1$ sous-intervalles de $[-\pi, \pi[$, et qu'avec une périodicité 12 au départ, ce nombre de sous-intervalles dépasse 12 quand $k$ dépasse 6.

# Références

Boudou, A. (1995) Mise en oeuvre de l'analyse en composantes principales d'une série stationnaire multidimensionnelle. *Pub. Inst. Stat. Univ.* Paris, XXXIX, fasc. 1, pp89-104
Boudou, A. and Dauxois, J. (1989) Analyses de mesures aléatoires, applications aux séries stationnaires. *C.R. Acad. Sci.* Paris, t.309, Série I, pp319-394
Brillinger, David R. (1975) *Time Series Data Analysis and Theory.* Rinehart and Winston, Inc Eds

# Une approche unificatrice pour l'estimation non-paramétrique des distributions de valeurs extrêmes multivariées

**Belkacem ABDOUS***

* Adresse pour correspondance :
Département de Médecine Sociale et Préventive
Université Laval, Québec

Exposé du 6 Mai 2002

## Résumé

L'exposé commencera par une présentation des principales caractérisations des distributions de valeurs extrêmes. On discutera ensuite les différentes techniques d'estimation non-paramétrique. On montrera que la majorité des techniques non-paramétriques existantes peuvent être unifiées et généralisées sous une classe d'estimateurs à noyau.

On montrera aussi que cette présentation permet l'élaboration de techniques de sélection optimale du paramètre de « lissage », commun à toutes les méthodes d'estimation des distributions de valeurs extrêmes.

On présentera aussi deux techniques permettant d'assurer que l'estimateur vérifie les propriétés d'une distribution de valeurs extrêmes.

On terminera l'exposé par une comparaison des principaux estimateurs.

## Sommaire des exposés des années précédentes

### Année 1999-2000

Sommaire de la publication # **LSP 2001-05**

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

**Année 2000-2001**
Sommaire de la publication # **LSP 2001-07**

**Journées de Statistique Fonctionnelle :**
**Toulouse 10-11 Juin 2002**
Sommaire de la publication # **LSP 2002-09**