Université Paul Sabatier

Habilitation à Diriger des Recherches

Modélisation Statistique pour Variables Aléatoires Fonctionnelles : Théorie et Application

Frédéric FERRATY

16 Juin 2003

Commission d'examen

B. CADRE	Univ. Montpellier II	Examinateur
W. GONZALEZ-MANTEIGA	Univ. Santiago, Espagne	Rapporteur/Examinateur
P. HALL	Univ. Camberra, Australie	Rapporteur
J. RAMSAY	Univ. McGill, Montreal, Canada	Rapporteur/Examinateur
Y. ROMAIN	Univ. Toulouse III	Examinateur
P. SARDA	Univ. Toulouse II	Examinateur
P. VIEU	Univ. Toulouse III	Directeur de recherche

Chercheur ou démarcheur
Gérant ou enseignant
Tel est le dilemme...

Remerciements

Je tiens à exprimer ma reconnaissance à Peter Hall pour avoir accepté d'être rapporteur de mes travaux et ainsi me faire bénéficier de son expertise internationalement reconnue.

Je remercie Wenceslao Gonzalez-Manteiga pour avoir également accepté d'être rapporteur de cette HDR. Sa vision à la fois globale et pointue des problèmes statistiques apporte un regard nouveau et pertinent, tant d'un point de vue théorique que pratique.

Quant au troisième rapporteur, Jim Ramsay, je tiens particulièrement à le remercier pour plusieurs raisons. D'une part, il est le spécialiste incontournable dans la modélisation et le traitement des données fonctionnelles. D'autre part, l'étendue et la qualité de ses travaux ont été pour moi un formidable catalyseur sur le plan scientifique. Pour toutes ces raisons, je tiens à le remercier très sincèrement.

Merci à Benoît Cadre pour l'intérêt qu'il a bien voulu accorder à mes recherches en acceptant de participer au jury.

Je voudrais aussi remercier Yves Romain pour sa participation au jury et l'intérêt qu'il porte à mes recherches au travers de nombreuses discussions fructueuses et en particulier concernant la Statistique des opérateurs, domaine dans lequel il excelle.

Je tiens également à remercier Pascal Sarda, non seulement pour sa participation au jury mais aussi pour les nombreuses collaborations scientifiques que nous avons menées ensemble. La richesse des échanges que nous avons eus, tant sur le plan scientifique qu'humain, a largement contribué à l'élaboration de cette HDR.

Quant à Philippe Vieu, je le remercie vivement d'avoir accepté d'être le Directeur de cette HDR. Mais au delà de cette fonction officielle, c'est avant tout le collaborateur que je voudrais remercier. En effet, les nombreuses et longues discussions que nous avons eues, la mise en commun de nos savoirs respectifs ainsi que la qualité permanente des rapports humains sont à l'origine d'une intense dynamique de recherche.

Je remercie aussi Hervé Cardot avec qui j'ai effectué de nombreux travaux. Ses remarquables qualités scientifiques associées à ses qualités humaines ont été un moteur permanent dans nos recherches. Merci aussi à André Mas pour ses collaborations fructueuses passées et prometteuses pour l'avenir.

Enfin, je remercie l'ensemble des collègues qui m'ont apporté aide, soutien, sympathie et amitié ainsi que tous ceux pour qui le mot science rime encore avec conscience. En particulier, je tiens à remercier Roger Lafosse pour sa vision de la recherche profondément humaniste, Françoise Michel pour sa gentillesse et son dévouement permanent ainsi qu'Elisabeth Lambert à qui je souhaite (en retard) une longue et heureuse retraite.

Plus généralement, c'est le groupe de travail STAPH que je tiens à remercier profondément car c'est à travers lui que nous avons pu créer une synergie autour de la Statistique Fonctionnelle. Le creuset de la science puise sa force dans la mise en commun des savoirs pour en retirer l'essence. C'est dans cet esprit que nous avons fondé STAPH, c'est toujours dans cet esprit que nous l'animons. Cette Habilitation en est une émanation, tant dans l'esprit que sur le plan scientifique. C'est tout naturellement que je dédie ces travaux à STAPH (*http ://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html*).

Table des matières

Chapitre 1

Résumé/Summary

1.1 Résumé

L'ensemble des travaux présentés traitent de la modélisation statistique pour des variables aléatoires à valeurs dans un espace $\mathcal{F}$ de dimension infinie. On parle alors de *variables aléatoires fonctionnelles* dont les observations fournissent des *données fonctionnelles* (*functional data*). Cet espace $\mathcal{F}$ peut être un espace de fonctions, d'opérateurs linéaires,.... Ainsi, la principale source des difficultés, tant d'un point de vue théorique que pratique, provient du fait que les observations de ce type de variables appartiennent à un espace infiniment dimensionné. Le fil conducteur des travaux réalisés réside dans l'apport systématique de solutions à ce problème de dimension infinie en mettant en place un cadre théorique suffisamment général. Les résultats obtenus puisent leur originalité et leur richesse dans la confrontation de plusieurs grands domaines de la statistique à savoir l'estimation fonctionnelle, la statistique des opérateurs, les variables fonctionnelles. De même, les outils nécessités pour les développements théoriques sont aussi de natures variées. En effet, ils relèvent de l'analyse fonctionnelle, de la théorie des opérateurs linéaires, de l'analyse numérique mais aussi d'outils probabilistes tels que la théorie des variables aléatoires hilbertiennes, les inégalités exponentielles pour des sommes de variables aléatoires, fonctionnelles ou non et plus récemment des probabilités de petites boules.

D'un point de vue pratique, de nombreux domaines d'applications alimentent ce thème moderne de la Statistique : chimiométrie, économétrie, agronomie, météorologie, biométrie,..., ce qui explique en partie l'engoue-

ment croissant de la communauté des statisticiens pour ce domaine. Ainsi, pour motiver et illustrer nos études, nous avons systématiquement mis en œuvre nos méthodes sur des données réelles.

Pour conclure, lors des diverses études menées, de nombreux problèmes ouverts sont apparus. Les travaux à venir découlant des nombreuses perspectives offertes par ce champ d'investigations ne manqueront pas de compléter et d'enrichir ceux déjà présents dans ce manuscrit.

1.2 Summary

Because this document is in french, a detailed summary have been produced. In addition, the works in relation with my PhD are not described here for concision reasons. The following presentation concerns only the studies performed after 1998.

General Presentation

The works presented in this document cover mainly and simultaneously three great parts of Statistics namely the Nonparametric Estimation, the Statistic of Operators and Functional Random Variables (or Functional Data). The tools used for theoretical developments are also of various kinds. Indeed, they come from functional analysis as well as form linear operators algebra or from numerical analysis. Of course, many probabilistic tools can be used : the theory arround hilbertian random variables, exponential inequalities for sums of (functional or not) random variables and recent small balls probabilities theory. In addition, practical aspects are omnipresent, which show the wide field of applications of such research topics.

First steps to the functional variables

Just after the defense of my PhD, I spent several months at McGill University (Montreal, Canada) in the laboratory of Jim Ramsay. At this time, the writing of the book entitled "Functional Data Analysis" by Ramsay and Silverman (1997) was in progress and I think that it was for me the starting point of my future investigations in these topics.

Functional variables and problematic

Before to go on, we recall that a functional random variable is a random variable taking its values in infinite-dimensional space $\mathcal{F}$. For instance, $\mathcal{F}$ can be a space of functions, linear operators,... Thus, the main difficulty, both in theory and practice, come from the fact that the observations of such variables are assumed to belong to an infinite-dimensional space. So, the guidelines and the main contribution of the jointed works consist in proposing a general theoretical framework able to override systematically the difficulties induced by this functional context. Moreover, applications and practical developments take an important place too.

Modelling and functional variables

A large part of my investigations concerns the regression models but others models have been studied in relation with time series, supervised classification of curves, distributions conditionally to a functional random variable. The common aim consists in estimating some mathematic objects (called generically operators) defined on the infinite-dimensional space $\mathcal{F}$. A model will be called parametric when the object to be estimated belongs to a set indexed by a finite number of parameters of $\mathcal{F}$ and nonparametric in the opposite case. Thus, according to this definition, the linear functional model can be viewed as a parametric model even if the target to be estimated is a functional object.

Two different approaches have been investigated almost simultaneously. The first (chronologically), parametric, focuses on the linear functional model. The second deals with nonparametric models.

Parametric approach : the linear functional model

Investigated jointly with Hervé Cardot and Pascal Sarda, we focused on the linear functional model which corresponds to a linear regression when the explanatory variable is a functional variable. First asymptotic properties have been stated for an empirical estimator based on functional principal components analysis.

Once the theoretical framework well defined, we emphasized on the necessity to obtain an estimator much smoother. Thus, we proposed two new estimators. The first one consists in projecting the previous empirical estimator onto a B-spline basis whereas the second one is a B-spline expansion obtained from a penalized least squares criterion. Finally, we compared these estimators, both theoritically and practically in various situations.

In addition, two statistics have been introduced in order to test the nullity of the functional coefficient of the functional linear model (in collaboration with André MAS). The first is approximated by a χ^2 whereas the second converges in distribution to a gaussian variable.

Nonparametric approach : models for functional variables

In collaboration with Philippe Vieu, we developped, just after the beginning of the previous parametric modelling, a nonparametric approach in order to build and study new models for functional variables.

In a first work, we focus on a nonparametric regression model when the regressor is a functional variable. The main difficulty is due to the so-called *curse of dimensionality*. Indeed, it is well known that the rate of convergence of some nonparametric estimator decreases when the dimension of the space in which the regressor takes its values increases. The principal originality of our work consists in solving this crucial *curse of dimensionality* by introducing a fractal type hypothesis acting on the functional variable. This fractal assumption allowed us to achieve first asymptotic results.

Once the theoretical framework well installed in the context of identical and independently distributed sample, we have been interested on modelling time series. To do this, the α-mixing have been considered and the previous results have been extended to this dependent case.

Afterthen, our kernel estimator introduced in the nonparametric regression model situation have been adapted in order to propose a nonparametric supervised classification for curves.

An another model seems very interesting to developp in this functional context namely the single index model. Our goal is to generalize this kind of models in the case when the explanatory variable is functional and the reulting model is called the single functional index model. First results have been stated for a fixed functional index but obviously, the most interesting is to obtain asymptotic properties for an unknown functional index (in progress).

The last field of investigations in the nonparametric context deals with the distribution conditionned by a functional variable. The cumulative density function, density function and its successive derivatives have been studied.

Finally, note that the small balls probabilities like $P(X \in d(x, X))$ appear in the asymptotic developments where X is a functional variable and d a metric or semi-metric (paper 12). In addition, as any process can be viewed as a functional variable, it is interesting to remark that the rate of convergence can be precised for some classical processes.

In conclusion

To conclude, I just recall two citations which are guidelines with respect to the presented works. The first is due to Bosq (1990) in a paper dealing with autoregressive hilbertian processes : *"These being nonparametric by themselves, it seems rather heavy to introduce a nonparametric model for observation lying in functional space ..."*. The second one can be found in the book of Ramsay et Silverman (1997) "Functional Data analysis" in the section entitled "Challenges for the future" : *"theoretical aspects of Functional Data analysis have not been researched in sufficient depth, and it is hoped that appropriate theoretical developments will feed back into advances in practical methodology"*. All the papers contained in this document give new (and sometimes first) answers to these thinkings. In addition, future investigations in this field will certainly contribute to complet and enrich the works already presented in this manuscrit.

Chapitre 2

Inroduction générale

Les travaux exposés dans ce document s'articulent simultanément autour de plusieurs grands domaines de la statistique à savoir l'estimation fonctionnelle, la statistique des opérateurs, les variables fonctionnelles. De même, les outils nécessités pour les développements théoriques sont de natures variées. En effet, ils relèvent de l'analyse fonctionnelle, de la théorie des opérateurs linéaires, de l'analyse numérique mais aussi d'outils probabilistes tels que la théorie des variables aléatoires hilbertiennes, les inégalités exponentielles pour des sommes de variables aléatoires, fonctionnelles ou non, sans oublier les résultats sur les probabilités de petites boules. Ainsi, les travaux réalisés puisent leur richesse et leur originalité dans un mariage parcimonieux de ces différents thèmes et outils. Enfin, notons l'omniprésence des aspects pratiques et des développements de procédures informatiques dans ces études, témoins des potentialités en terme d'applications de ce vaste champ de recherche.

Juste après la soutenance de ma thèse[1], j'ai effectué un post-doc à l'Université McGill de Montreal (Canada) où j'ai rencontré Jim Ramsay, lequel

[1] Bien qu'éloigné des travaux présents dans ce document, mon travail de thèse fut réalisé dans le même esprit en ce sens qu'il était basé sur un mariage de deux domaines de la Statistique relativement distants. L'un, concernait ce que l'on appelle usuellement l'analyse factorielle multidimensionnelle alors que l'autre relevait de l'estimation nonparamétrique. L'idée directrice de ce travail était de rechercher une structure commune en combinant des méthodes de réduction de dimension (analyse factorielle) avec une étape de lissage. En effet, dans certaines situations, cette réduction de dimension peut s'avérer trop grossière. D'où l'idée d'y incorporer une étape d'estimation nonparamétrique (lissage) afin de rendre plus souple la réduction de dimension. Cependant, les résultats théoriques étant difficiles à atteindre dans ce genre d'études, les aspects pratiques avaient pris une place prépondérante.

finissait à l'époque la rédaction du livre intitulé "Functional Data Analysis". Je crois bien que c'est à cette époque que j'ai réalisé l'importance de ce thème de recherche, où seuls les aspects pratiques avaient commencé à émerger alors que les approches théoriques restaient marginales. Finalement, ce sera lors de discussions fructueuses avec Hervé Cardot, Pascal Sarda et Philippe Vieu que nous avons conjointement décidé d'investir ce thème de recherche et d'aborder la modélisation statistique des variables fonctionnelles.

Rappelons qu'une variable aléatoire fonctionnelle est tout simplement une variable aléatoire à valeurs dans un espace de dimension infinie $\mathcal{F}$. Par exemple, cet espace $\mathcal{F}$ peut être un espace de fonctions, d'opérateurs linéaires,.... Selon la terminologie en vigueur dans la littérature, on parle aussi bien de variables aléatoires fonctionnelles que de données fonctionnelles, ce qui englobe notamment tout ce qui concerne l'analyse statistique de courbes. La principale source des difficultés, tant d'un point de vue théorique que pratique, provient du fait que les observations de ce type de variables sont supposées appartenir à un espace infiniment dimensionné. Le fil conducteur des travaux présents dans ce document réside dans l'apport systématique de solutions à ce problème de dimension infinie en mettant en place un cadre théorique suffisamment général. Par ailleurs, vu les nombreux domaines d'applications concernés par ce champ d'investigations, les aspects pratiques n'ont jamais été négligés. Bien que les modèles de régression tiennent une place prépondérante, d'autres types de modélisation ont été abordés, notamment ceux concernant les séries temporelles ainsi que la classification supervisée de courbes. Par ailleurs, on va être amené à estimer des objets mathématiques (appelés génériquement opérateurs) définis sur l'espace $\mathcal{F}$ de dimension infinie et à valeurs dans $\mathbb{R}$. La nature du problème statistique posé est doublement fonctionnelle ; elle est liée aussi bien à l'objet que l'on souhaite estimer qu'aux variables que l'on observe. On parlera alors de modèle paramétrique lorsque l'objet à estimer appartient à un ensemble indexable par un nombre fini de paramètres appartenant à $\mathcal{F}$ et de modèle nonparamétrique dans le cas contraire. Ainsi, comme on le verra par la suite, selon cette définition le modèle linéaire fonctionnel peut être considéré comme un modèle paramétrique bien que l'objet à estimer soit de nature fonctionnelle. Deux approches différentes ont été menées pratiquement de front. La première (d'un point de vue chronologique), paramétrique, s'est concentrée autour du modèle linéaire fonctionnel. La seconde aborde la modélisation de variables fonctionnelles de façon purement nonparamétrique.

Le Chapitre 2 précisera la problématique générale dans laquelle nous nous plaçons. Des exemples concrets seront présentés afin de souligner les nombreuses potentialités d'un point de vue applicatif. Puis, l'ensemble de ces travaux seront positionnés dans la littérature. Sans prétendre à un état de l'art de ce vaste domaine, ce sera néanmoins l'occasion de donner un éventail des avancées récentes autour des variables aléatoires fonctionnelles. Nous profiterons aussi de ce chapitre pour donner quelques définitions élémentaires, notamment concernant la terminologie paramétrique et nonparamétrique dans ce cadre fonctionnel. Enfin, une liste des travaux présentés dans ce manuscrit achèvera ce chapitre.

Une troisième partie sera consacrée aux développements réalisés autour du modèle fonctionnel de régression linéaire. Il s'agit de régresser une v.a.r. Y sur une v.a.f. X, la liaison entre Y et X se faisant au travers d'un opérateur linéaire. Or, il est connu que la régression linéaire multivariée est défaillante lorsque le nombre de variables explicatives est important comparativement à la taille de l'échantillon considéré ou bien lorsqu'il existe des colinéarités parmi les régresseurs. Ces deux situations pathologiques se produisent simultanément quand les variables explicatives correspondent à une discrétisation fine d'une variable aléatoire fonctionnelle. Pour pallier à ces problèmes, plusieurs méthodes heuristiques telles que la Ridge Regression, PCR (Principal Component Regression) ou encore PLSR (Partial Least Squares Regression) ont été proposées dans la littérature. Cependant, l'absence d'étude asymptotique dans ce domaine ainsi que la déficience dans certaines situations des estimateurs existants nous a conduit à mettre en place un cadre théorique afin d'exhiber les propriétés d'estimateurs que nous avons introduits. Plus récemment, ces travaux ont été complétés par la mise en place de tests fonctionnels portant sur un tel modèle.

Quant à la quatrième partie, elle se focalisera sur les modèles nonparamétriques pour variables aléatoires fonctionnelles. Dans un premier temps, on considère le modèle de régression nonparamétrique fonctionnel pour lequel la variable explicative est fonctionnelle alors que la variable réponse est réelle. Le principal apport fut d'avoir solutionné le problème de la dégradation de la dimension (*Curse of dimensionality*) bien connu en statistique nonparamétrique. D'un point de vue général, losqu'on estime nonparamétriquement un objet mathématique fonction de p variables aléatoires, on sait que la vitesse de convergence d'un tel estimateur décroit quand p augmente. Maintenant, considérer une variable aléatoire fonctionnelle revient à prendre $p = \infty$.

Cette étude nous permettra de mettre en place un cadre théorique et d'obtenir des premiers résultats asymptotiques. Dans un second temps, nous étendrons ces résultats aux séries temporelles. D'autres développements seront abordés comme la classification supervisée de courbes, la caractérisation de distributions conditionnellement à une variable fonctionnelle, ainsi qu'un premier modèle de réduction de dimension dans ce cadre doublement fonctionnel à savoir le modèle à indice fonctionnel simple.

Une cinquième partie proposera quelques exemples d'applications qui soulignent l'intérêt pratique de ces méthodes pour variables fonctionnelles.

Enfin, une dernière partie fera état de la multitude de problèmes ouverts résultant de la synthèse de l'ensemble des travaux présentés dans ce document. La profusion ainsi que l'éventail des perspectives offertes alimenteront très certainement de nombreux travaux, tant sur un plan théorique que pratique, et cela pour plusieurs années.

Chapitre 3

Variables fonctionnelles : problématique et positionnement des travaux dans la littérature

L'objectif de ce chapitre est triple. Dans un premier temps, quelques exemples pratiques permettent d'illustrer notre problématique alors qu'une terminologie est introduite au travers de défintions élémentaires. Dans un second temps, des aspects bibliographiques sont développés afin de situer nos travaux dans la littérature. Il s'agit de fournir à la fois une idée du cadre de travail dans lequel on se place mais aussi quelques mots clés permettant une meilleure appréhension de la littérature. Enfin, la listes des travaux présentés dans ce document terminera cette partie.

3.1 Problématiques : exemples et définitions élémentaires

Les progrès de l'outil informatique tant en capacité de mémoire qu'en calcul permettent d'enregistrer des données de plus en plus volumineuses. Ainsi, un très grand nombre de variables peuvent être observées pour l'étude d'un même phénomène. C'est notamment le cas lorsqu'on dispose d'une famille de variables $\{X(\theta)\}_{\theta\in\Theta}$ indexée par un paramètre θ variant dans un espace "continu" Θ (par exemples, $\Theta \subset \mathbb{R}$ ou $\Theta \subset \mathbb{R}^p$ ou encore $\Theta \subset H$ avec H un espace hilbertien,...). Bien évidemment, il est techniquement impossible de mesurer $X(\theta)$ pour tout θ dans Θ. Néanmoins, on dispose d'une discrétisation suffisamment fine $\{\theta_i\}_{i=1,\dots,I}$ de Θ pour pouvoir considérer que le comportement de $\{X(\theta)\}_{\theta\in\Theta}$ est "proche" de celui de $\{X(\theta_i)\}_{i=1,\dots,I}$. Une telle famille $X = \{X(\theta)\}_{\theta\in\Theta}$ de variables aléatoires sera appelée variable aléatoire fonctionnelle. Le terminologie "fonctionnel" est assez naturelle car une simple observation d'une telle variable fournit un objet appartenant à un espace fonctionnel (de dimension infinie). Le premier exemple concerne des données spectrométriques couramment utilisées en chimie quantitative. On s'intéresse à un échantillon de 215 morceaux de viandes finement hachées contenant chacun un certain taux de matière grasse (lipide). Pour chaque morceau, on observe le spectre dans le proche infrarouge pour 100 longueurs d'onde $\{\lambda_j\}_{j=1,\dots,100}$ réparties entre 850 et 1050 nanomètres avec un pas constant. On observe donc pour chaque morceau i, une famille "discrète" $X_i^d = \{X_i(\lambda_j)\}_{j=1,\dots,100}$ que l'on peut considérer comme une version discrétisée de la variable fonctionnelle $X_i = \{X_i(\lambda)\}_{\lambda\in[850;1050]}$. On parle dans ce cas de courbe spectrométrique. Leurgans et *al.* (1993) conforte cette idée en remarquant que "the spectra observed are to all intents and purposes functional observations" alors que la représentation des 215 spectres ci-dessous fait apparaitre clairement l'aspect fonctionnel :

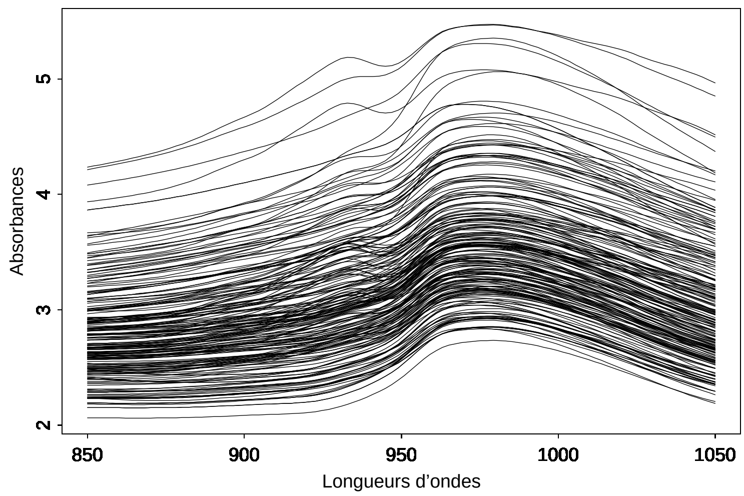

FIG. 3.1 – Courbes spectrométriques

Problématique : étant donné une courbe spectrométrique, peut-on estimer le taux de lipide du morceau de viande correspondant ? Pour pouvoir répondre à cette question, on connait le taux de lipide Y_i contenu dans le morceau i, lequel a été obtenu à partir d'une analyse chimique fiable mais plus coûteuse en temps et économiquement que le serait le procédé spectrométrique. Ainsi, on va être amené à modéliser pour $i = 1, \ldots, 215$, le lien entre la variable réelle Y_i et la variable "fonctionnelle" X_i.

Le second exemple est tiré du domaine agronomique. Il s'agit de mieux comprendre les interactions entre le rendement de blé et les variations climatiques survenues durant la culture. Dans ce but, on dispose de $n = 198$ parcelles de blé ; pour chaque parcelle i, en plus du rendement Y_i, on a les températures T_i^d et précipitations P_i^d cumulées journalières durant $J = 309$ jours (d'Octobre à début Août) : $T_i^d = \{T_i(t_j)\}_{j=1,\ldots,J}$ et $P_i^d = \{P_i(t_j)\}_{j=1,\ldots,J}$. Le graphique suivant représente la version "continue" $\{T_i = \{T_i(t) \,; t \in [0; 309]\}\}_{i=1,\ldots,n}$ (resp. $\{P_i = \{P_i(t) \,; t \in [0; 309]\}\}_{i=1,\ldots,n}$) de $\{T_i^d\}_{i=1,\ldots,n}$ (resp. $\{P_i^d\}_{i=1,\ldots,n}$) : Une

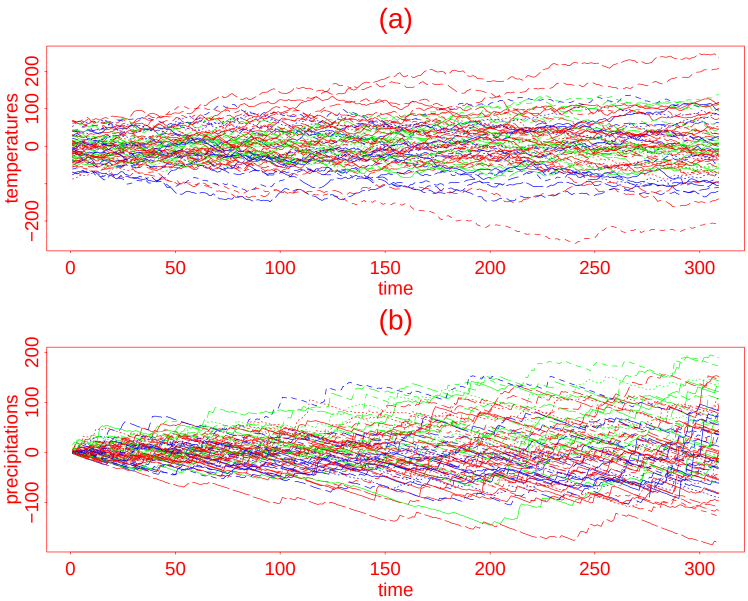

Fig. 3.2 – Courbes de pluviométrie et de température ; relevés journaliers cumulés centrés

fois de plus, on s'intéresse pour $i = 1, \ldots, n$ à la modélisation du lien entre la variable réelle Y_i (rendement de blé) et le couple de variables fonctionnelles

(P_i, T_i) (pluviométrie et température).

Un troisième exemple issu du domaine de l'économétrie consiste en l'observation des fluctuations d'un indice boursier en fonction du temps : Il

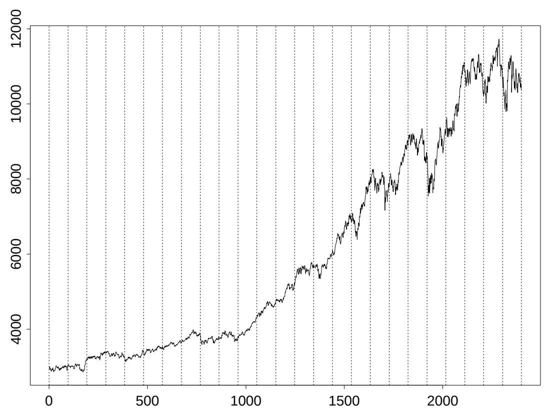

FIG. 3.3 – Évolution du downjones

s'agit typiquement d'une série temporelle. Les traits verticaux symbolisent une subdivision selon l'axe horizontal, laquelle permet d'obtenir l'échantillon de courbes suivant :

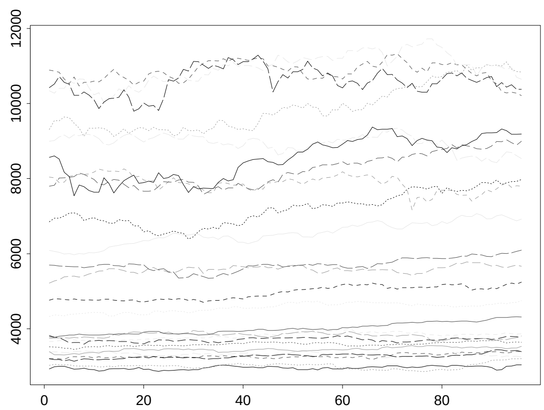

FIG. 3.4 – Échantillon de courbes obtenu après découpage de la série initiale

Autrement dit, ce découpage en sous-intervalles nous ramène à plusieurs observations d'une variable fonctionnelle. Nous verrons par la suite les différentes problématiques qui peuvent être proposées dans un tel contexte.

Une dernière illustration de variable fonctionnelle concerne la reconnaissance vocale. Plus précisément, on se place dans un cadre de classification supervisée de cinq catégories de phonèmes. Les données sont des log-périodogrammes correspondant à l'enregistrement de locuteurs sur une durée de 32 millisecondes. Cette étude s'intéresse à la discrimination des cinq phonèmes suivantes : "sh" comme dans "she", "dcl" comme dans "dark", "iy" comme dans "she", "aa" comme dans "dark" and "ao" comme dans "water". Pour chaque type de phonème, on dispose de 400 enregistrements obtenus à une fréquence d'échantillonnage de 16-kHz ; seules les 150 premières fréquences sont retenues. Le graphique suivant représente dix log-périodogrammes pour chaque type de phonème :

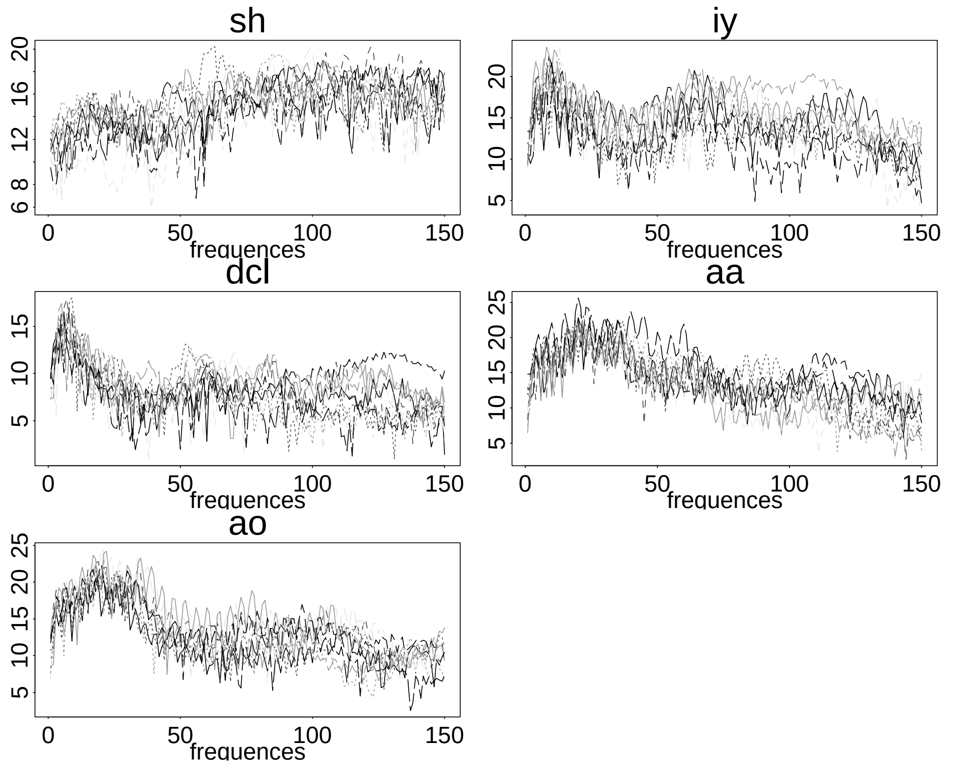

FIG. 3.5 – Courbes phonétiques

La problématique est simple : à partir d'un log-périodogramme, comment peut-on lui attribuer la phonème correspondante ?
Une fois la réalité des variables fonctionnelles exhibées, on peut s'intéresser aux aspects de modélisation les concernant. Dans ce but, nous donnons quelques définitions permettant de fixer un vocabulaire une fois pour toute.

Définition 3.1.1 *On appelle* **variable aléatoire fonctionnelle (v.a.f.)***, toute application mesurable définie sur un espace probabilisé* $(\Omega, \mathcal{A}, P)$ *et à valeurs dans un espace probabilisable* $(E, \mathcal{B}_E)$ *où* E *est un espace fonctionnel de dimension infinie et* $\mathcal{B}_E$ *la tribu des boréliens de* E*.*

Définition 3.1.2 *On appelle* **modèle fonctionnel***, tout modèle prenant en compte au moins une v.a.f.*

Ayant défini les v.a.f., on peut aller un peu plus loin dans les aspects de modélisation. En particulier, on va distinguer, comme dans le cadre de variables non fonctionnelles, deux grandes familles de modèles : les modèles dits paramétriques et ceux dits nonparamétriques. De façon générale, les modèles étudiés dans ce document ne fournissent pas les lois des variables mises en jeu. En revanche, ils exhibent une fonction Φ appartenant à l'ensemble $\mathcal{F}_E^{E'}$ des fonctions définie sur l'espace fonctionnel E et à valeurs dans E'. L'estimation d'un tel objet mathématique sera rendue possible grâce à une condition de la forme

$$\Phi \in \mathcal{C} \tag{3.1}$$

où $\mathcal{C}$ est un sous-ensemble de $\mathcal{F}_E^{E'}$.

Définition 3.1.3
Un modèle fonctionnel est dit **paramétrique** *si* $\mathcal{C}$ *est indexable par un nombre fini de paramètres appartenant à* E*.*
Un modèle fonctionnel est dit **nonparamétrique** *dans le cas contraire.*

Bien entendu, il existe d'autres modèles fonctionnels à mi-chemin entre le paramétrique et le nonparamétrique. C'est notamment le cas du modèle à indice fonctionnel simple que l'on pourrait qualifier de modèle fonctionnel semiparamétrique et que nous présentons à la fin de ce paragraphe.
Dans tout ce qui suit, X désigne une v.a.f. et Y une variable aléatoire quelconque. Soit $(X_i, Y_i)_{i=1,\dots,n}$ n couples identiquement distribués selon la loi de (X, Y).

Définition 3.1.4
De façon générique, on appelle **modèle fonctionnel de régression** *tout modèle s'écrivant sous la forme*

$$Y_i \;=\; \Phi(X_i) + \varepsilon_i, \tag{3.2}$$

où les ε_i's sont n variables aléatoires centrées. X est appelé indifférement "variable explicative fonctionnelle" ou "régresseur fonctionnel" alors que Y est la "variable à expliquer" ou encore la "réponse".

Pour fixer les idées, voici deux exemples de modèles fonctionnels de régression que l'on verra plus en détail par la suite. Soit Y une v.a.r. On peut alors définir un modèle fonctionnel de la façon suivante :

$$Y \;=\; \Phi(X) + \varepsilon, \tag{3.3}$$

avec Φ un opérateur linéaire continu et ε une v.a.r. centrée. On parle alors de **modèle fonctionnel de régression linéaire**. En reprenant les notation précédentes, $\mathcal{C} = \{$ensemble des opérateurs linéaires continus de E dans $\mathbb{R}\}$. Or, d'après le théorème de représentation de Riesz, il existe un unique élément ϕ de E tel que $\Phi(X) =< \phi, X >$ où $< .,. >$ désigne le produit scalaire dans E. Ainsi, $\mathcal{C} = \{\phi \in E\}$. Autrement dit, $\mathcal{C}$ est indexable par un seul paramètre appartenant à E. Ce modèle appartient à la famille des modèles fonctionnels paramétriques

Un autre type de modèle fonctionnel de régression peut être donné de la manière suivante :

$$Y \;=\; \Phi(X) + \varepsilon, \tag{3.4}$$

avec Φ tel que

$$\exists C > 0,\ \exists \beta > 0,\ \forall (x_1, x_2) \in E \times E,\ |\Phi(x_1) - \Phi(x_2)| \leq C d(x_1, x_2)^\beta,$$

ε une v.a.r. centrée et où d est par exemple une semi-métrique sur E. Dans ce cas,

$$\mathcal{C} = \{\Phi \in \mathcal{F}_E^{\mathbb{R}} \,/\, \exists C > 0,\ \exists \beta > 0,\ \forall (x_1, x_2) \in E \times E,\ |\Phi(x_1) - \Phi(x_2)| \leq C d(x_1, x_2)^\beta\}.$$

Il est clair qu'une telle famille n'est pas indexable par un nombre fini de paramètre de E. On parle alors de **modèle fonctionnel nonparamétrique de régression**.

Un dernier exemple concerne le **modèle à indice fonctionnel simple qui s'écrit** :

$$Y \;=\; \Phi_2(< \phi_1, X >) + \varepsilon, \tag{3.5}$$

que l'on peut formuler de manière équivalente comme suit :

$$Y \;=\; \Phi_2 \circ \Phi_1(X) + \varepsilon, \tag{3.6}$$

où, Φ_1 est un opérateur linéaire continu et Φ_2 tel que

$$\exists C > 0,\ \exists \beta > 0,\ \forall (x_1, x_2) \in E \times E,\ |\Phi_2(x_1) - \Phi_2(x_2)| \leq C d(x_1, x_2)^{\beta}.$$

Ce modèle fonctionnel se situe entre les modèles fonctionnels paramétriques et nonparamétriques ; on parle alors de modèle fonctionnel semiparamétrique.

3.2 Tour d'horizon sur les v.a.f. et positionnement dans la littérature

Sans prétendre à une quelconque exhaustivité, ce paragraphe propose un tour d'horizon sur la littérature concernant les v.a.f. Les travaux qui font l'objet de ce document ont été volontairement dissociés du texte ; ils apparaissent systématiquement en fin de paragraphe pour des raisons de lisibilité. Par ailleurs, un découpage thématique est réalisé afin de mieux structurer la bibliographie. Enfin, historiquement, les premières approches statistiques concernant les v.a.f. ont été réalisées dans le cadre de l'analyse factorielle (notamment l'analyse en composantes principales fonctionnelle). Il est donc naturel d'inclure dans la bibliographie ces aspects même si nos travaux ne s'y inscrivent pas directement.

3.2.1 Bref historique ; analyses factorielles et v.a.f.

Les premiers développements statistiques opérant sur des variables fonctionnelles concernaient les analyses factorielles. Ils reposent essentiellement sur la décomposition de Karhunen-Loeve (1945) permettant une représentation optimale au sens de la variance de la projection d'une fonction aléatoire sur un espace de dimension finie. Il faut remonter jusqu'au début des années soixante pour trouver les premiers balbutiements. En effet, ce sont des climatologues (Obhukov, 1960, Holmström, 1961, Buell, 1971) qui, en s'intéressant à des champs de pression, ont proposé les premières analyses factorielles de variables fonctionnelles. Peu à peu, d'autres disciplines scientifiques se sont intéressées à des observations sous forme de trajectoires : Deville (1974) dans le domaine de l'économie, Molenaar et Boomsma (1987) puis Kirkpatrick et Heckman (1989) en génétique. Parallèlement, Dauxois et Pousse (1976), Dauxois et *al.* (1982) puis Antoniadis et Beder (1989) fournirent des études asymptotiques portant sur l'estimation des éléments spectraux d'opérateurs issus d'analyses factorielles de fonctions aléatoires. Saporta (1981) proposa une étude synthétique de méthodes exploratoires d'analyse de processus. Plus récemment, Leurgans et *al.* (1993) proposèrent une étude à la fois théorique et pratique de l'analyse canonique de deux v.a.f. alors que de nombreux auteurs ont développé des méthodes autour de l'analyse en composantes principales de v.a.f. (Ramsay et Dalzell, 1991, Pezzulli et Silverman, 1993, Silverman, 1995 et 1996, Besse et *al.*, 1997, Locantore et *al.*, 1999, James et *al.*, 2000, Preda, 2000). Enfin, la plupart des développements asymptotiques sont fondés sur la Statistique Opératorielle (voir pour les travaux les plus récents Dauxois et *al.*, 2001, Boudou et Romain, 2002, Romain, 2002) dont Romain (2000) fournit une excellente synthèse.

3.2.2 Bibliographie récente et positionnement de mes travaux

Modèles fonctionnels de régression

Les modèles fonctionnels de régression tiennent une place privilégiée dans nos travaux. En effet, nos investigations se sont d'abord focalisées sur ce type de modélisation. Il convient alors de préciser la terminologie inhérente à ces modèles. Dans ce but, considérons $(X_i, Y_i)_{i=1,\dots,n}$, n couples identiquement distribués selon la loi de (X, Y) avec X v.a.f., et reprenons la définition d'un modèle fonctionnel de régression (3.2). On peut alors distinguer diverses situations :

i) Y est une réponse scalaire ou fonctionnelle,

ii) l'échantillon est formé de couples indépendamment distribués (Indep.) ou le contraire (Dep.),

iii) modèle fonctionnel paramétrique ou nonparamétrique.

Le tableau ci-dessous récapitule ces diverses situations :

	Modèle fonctionnel			
	Paramétrique		Nonparamétrique	
Échant. / Réponse	Indep.	Dep.	Indep.	Dep.
Scalaire	(1)		(4)	(5)
Fonctionnelle	(2)	(3)		(6)

Tab. 3.1 – Synoptique des diverses situations

Nous précisons maintenant point par point la littérature inhérente à ces différents contextes identifiés par la numération du Tableau 3.1 :

Situation (1)

De nombreux statisticiens ont développé des applications adaptées aux v.a.f. lorsque la variable réponse est scalaire (réelle). Seulement, pendant très longtemps, la nature fonctionnelle de ces données n'a été que partiellement prise en compte. En effet, les premiers travaux se sont focalisés directement sur la version discrétisée des v.a.f. observées, ce qui revient à considérer un nombre fini de régresseurs. Frank et Friedman (1993) proposent une importante bibliographie liée à des applications dans le domaine de la chimie quantitative. Qu'il s'agisse de *ridge regression* (Hoerl et Kennard, 1970), de régression *partial least squares* (Wold, 1975, Martens et Naes, 1989) ou de régression sur composantes principales (Massy, 1965), toutes ces méthodes ont pour but d'adapter la régression linéaire classique (*ordinary least squares regression*) lorsque le nombre de régresseurs est très important devant le nombre d'individus ou bien encore lorsque les régresseurs sont très corrélés, ces deux cas pathologiques pouvant se produire simultanément. Il faut attendre la discussion de Hastie et Mallows en 1993 pour voir apparaître la version fonctionnelle du modèle de régression linéaire classique c'est-à-dire le modèle (3.3). Un premier travail synthétique dans ce cadre est fourni par Preda (1999), lequel propose une extension de la méthode

PLS au cadre fonctionnel.
Contributions : Mise en place du cadre théorique permettant l'obtention de premiers résultats asymptotiques dans le modèle fonctionnel de régression linéaire (Cardot, Ferraty et Sarda, 1999) ; introduction puis études des propriétés asymptotiques d'estimateurs "lisses" (Cardot, Ferraty et Sarda, 2003) ; mise en place de tests (Cardot, Ferraty, Mas et Sarda, 2003).

Situation (2)
Les premiers travaux théoriques dans les modèles fonctionnels de régression sont dus à Bosq (1991) dans le cas de v.a.f. dépendantes (processus autorégressifs hilbertiens). Il s'agit de modèles fonctionnels paramétriques (au sens défini dans le paragraphe précédent) où la variable réponse est aussi fonctionnelle. Depuis, de nombreux développements asymptotiques concernant ces processus linéaires ont été realisés (Bosq, 96 et 99, Merlevède et *al.*, 97, Mas, 99, Pumo, 99, Antoniadis, 2002, Damon and Guillas, 2002).

Situation (3)
Le seul papier concernant ce type de modèle a été réalisé par Cuevas et *al.* (2001). Ils ont fourni des premiers résultats théoriques dans le cadre de la régression linéaire fonctionnelle d'une v.a.f. sur une fonction déterministe ("fixed design").

Situation (4)
Aucun travail n'existant sur ce sujet, nous avons fourni les premiers résultats, tant d'un point de vue théorique que pratique.
Contributions : Estimation et propriétés théoriques de la régression d'une variable réelle sur une v.a.f. avec applications (Ferraty et Vieu, 2000, Ferraty et Vieu, 2002).

Situation (5)
Ici aussi, nous avons produit les premiers développements théoriques et pratiques dans cette situation. Depuis, nos méthodes ont été appliquées à des données de pollution dans un travail récent dû à Aneiros et *al.* (2002).
Contributions : prédiction et v.a.f. dépendantes ; théorie

ET APPLICATION (FERRATY, GOIA ET VIEU, 2002A, FERRATY ET VIEU, 2003B).

Situation (6)

Très peu de littérature existe sur ce domaine. Toutefois, on peut citer Bosq et Delecroix (1985) qui étudièrent une prédiction nonparamétrique d'une v.a.f. dans le cadre de processus markovien strictement stationnaire. Dans le contexte de v.a.f. dépendantes, Besse et *al.* (2000) ont mis en oeuvre une méthode nonparamétrique de prédiction (réponse fonctionnelle) sur des données climatiques.

Modèles fonctionnels semi-paramétriques

À notre connaissance, la seule étude rentrant dans ce sujet est due à Ferré et *al.* (2001) dans le cadre d'une extension de la méthode SIR aux v.a.f. et dont une partie des outils utilisés est similaire à l'approche théorique de Cardot, Ferraty et Sarda (1999).
CONTRIBUTIONS : MODÈLE À INDICE FONCTIONNEL SIMPLE (FERRATY, PEUCH ET VIEU, 2003).

Classification de courbes

Différents auteurs se sont intéressés au cas où on dispose d'un échantillon d'un couple (X, Y) avec X v.a.f. et Y qualitative. On parle alors aussi bien de discrimination que de classification supervisée de courbes. Hastie et *al.* (1995) proposèrent une version pénalisée de l'analyse discriminante classique permettant de discriminer linéairement des courbes alors que Marx et Eilers (1999) étudièrent une version fonctionnelle du modèle linéaire généralisé. Quant à Hall et *al.* (2001), ils proposèrent une méthode nonparamétrique de discrimination de courbes (réponse qualitative). Dans un contexte purement de classification (on n'observe pas Y), citons le travail d'Abraham et *al.* (2002).
CONTRIBUTIONS : CLASSIFICATION SUPERVISÉE (DISCRIMINITION) DE COURBES (FERRATY ET VIEU, 2003A)

Distributions et v.a.f.

Il s'agit ici de travaux en lien avec l'estimation de la densité de distribution de v.a.f. Des premiers travaux s'inscrivent dans cette optique (Geffroy, 1974,

Jacob and Oliveira, 1995). Plus récemment, Gasser et *al.* (1998) puis Hall et Heckman (2002) se sont intéressés de manière heuristique à l'estimation nonparamétrique du mode d'une distribution de v.a.f. Levallois (1998) puis Berlinet et Levallois (2001) donnent en préambule des notions générales (en particulier la dimension fractale) s'appliquant aux mesures de v.a.f. Mais il faut attendre les travaux de Niang (2002a, 2002b) pour avoir les premiers résultats asymptotiques sur l'estimation de la densité de v.a.f.
<u>Contributions :</u> Caractérisations d'une distrbution conditionnelle à une v.a.f. (Ferraty, Laksaci et Vieu, 2003, soumis) ; courbes modales (Ferraty, Niang et Vieu, en cours).

Extensions diverses pour v.a.f.

Bien entendu, tous les travaux concernant les v.a.f. ne s'inscrivent pas forcément dans le découpage thématique ci-dessous. C'est notamment le cas de l'analyse de variance (ANOVA) fonctionnelle (Luo, 1998, Luo et *al.*, 1998), des "times-varying coefficient models" (Hoover et *al.*, 1998, James et *al.*, 2000), des problèmes de recalage de courbes (Kneip et Gasser, 1992, Ramsay et Li, 1996), des modèles à effets mixtes (Rice et Silverman, 1991, Boularan et *al.*, 1995, Brumback et Rice, 1998, Wang, 1998, Nunez-Anton et *al.*, 1999, Rice et Wu, 2001). Enfin, citons les travaux filtrant les données au moyen d'opérateurs différentiels qui ont prouvé leur efficacité dans de nombreux exemples (Ramsay, 1996, Ramsay et *al.*, 1994, Ramsay et *al.*, 1995).

Ouvrages de synthèse

Pour finir ce tour d'horizon, citons deux monographies incontournables dans le domaines des v.a.f. : Ramsay et Silverman (1997, 2002) qui, en plus d'une bibliographie bien étoffée, exposent un large éventail des méthodes statistiques capables de traiter les v.a.f. et Bosq (2000) qui s'est focalisé sur la modélisation de v.a.f. dépendantes en donnant de nombreux développements théoriques.
<u>Contributions :</u> Modèles de régression pour variables aléatoires réelles, vectorielles et fonctionnelles (Ferraty, Núñez Antón et Vieu, 2001, Ferraty, Goia et Vieu, 2002), Statistique nonparamétrique fonctionnelle (Ferraty et Vieu, 2003).

3.3 Liste des travaux

3.3.1 Articles

27) F. Ferraty, A. Peuch et P. Vieu. Modèle à indice fonctionnel simple. *À paraître dans C. R. Acad. Sci. Paris.*

26) H. Cardot, F. Ferraty et P. Sarda (2003). Spline Estimators for the Functional Linear Model. *À paraître dans Statistica Sinica.*

25) F. Ferraty and P. Vieu (2003). Nonparametric Models for Functional Data with Application in Regression, Time Series and Curve Discrimination. *Accepté pour publication dans J. of Nonparametric Statist.*.

24) F. Ferraty and P. Vieu (2003). Curves discrimination : a nonparametric functional approach. *À paraître dans Computational Statistics & Data Analysis.*

23) H. Cardot, F. Ferraty, A. Mas et P. Sarda (2003). Test of nullity in the Functional Linear Model. *Scand. J. Stat.*, **30**, 241-255.

22) F. Ferraty, A. Goia, P. Vieu (2002). Functional Nonparametric Model for Time Series : a Fractal Approach for Dimension Reduction. *TEST*, **11**, 317-344.

21) F. Ferraty, P. Vieu (2002). The Functional Nonparametric Model and Application to Spectrometric Data. *Computational Statistics*, **17**, 545-564.

20) F. Ferraty, A. Goia, P. Vieu (2002). Régression non-paramétrique pour des variables aléatoires fonctionnelles mélageantes. *C. R. Acad. Sci. Paris*, Ser. I 334, 217-220.

19) H. Cardot, F. Ferraty, P. Sarda (2000). Étude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. *C. R. Acad. Sci. Paris*, t. 330, Série I, 501-504.

18) F. Ferraty, P. Vieu (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *C. R. Acad. Sci. Paris*, t. 330, Série I, 139-142.

17) H. Cardot, F. Ferraty, P. Sarda (1999). Linear Functional Model. *Statist. & Prob. Letters*, **45**, 11-22.

16) P. C. Besse, H. Cardot, F. Ferraty (1997). Simultaneous Nonparametric Regressions of Unbalanced Longitudinal Data, *Computational Statistics & Data Analysis*, **24**, 255-270.

15) F. Ferraty (1997). Estimations de Transformations Optimales en ACP Curvilinéaire, *Revue de Statistique Appliquée*, **XLV** (1), 5-39.
14) P. C. Besse, F. Ferraty (1995). Curvilinear Fixed Effect Model, *Computational Statistics*, **10**, 339-351.

3.3.2 Article soumis

13) F. Ferraty, A. Laksaci and P. Vieu. Estimating some characteristics of the conditional distribution in nonparametric functional models. *Soumis pour publication.*

3.3.3 Ouvrages

12) F. Ferraty et P. Vieu. Rédaction (**sur invitation**) d'un chapitre "Functional nonparametric statistics : a double infinite dimensional framework" pour un ouvrage collectif intitulé "Recent Advances and Trends in Nonparametric Statistics" (en cours).
11) F. Ferraty, A. Goia, P. Vieu (2002). Statistica funzionale ; Modelli di regressione non-parametrici. Ed. Franco Angeli, Collana : Economia e Statistica, Milano.
10) F. Ferraty et P. Vieu (2002). Statistique Fonctionnelle : Modèles nonparamétriques de régression. Notes de cours de DEA 2002/2003 (accessible à l'adresse *http ://www.lsp.ups-tlse.fr/Fp/Ferraty/RNP.ps*).
9) F. Ferraty, V. Nunez-Anton, Vieu P. (2001). Regressión No Paramétrica : Desde la Dimensión Uno Hasta la Dimensión Infinita. Servicio Editorial de la Universidad del Pais Vasco, Bilbao.

3.3.4 Publications internes

8) Cardot H., Ferraty F. et Sarda P. (1999). Spline Estimators for the Functional Linear Model : Consistency, Application and Splus Implementation. *Rapport UBIA, 1999/1, INRA, Toulouse.*
7) Cardot H., Ferraty F. et Sarda P. (1998). Modèle linéaire fonctionnel. *Publication du laboratoire de Statistique et Probabilités, 04-98, Toulouse.*
6) Ferraty F. (1995). Modèle curvilinéaire pour l'analyse discriminante. *Publication du laboratoire de Statistique et Probabilités.*

5) Besse P. et Ferraty F. (1994). Modèle curvilinéaire à effet fixe. *Publication du laboratoire de Statistique et Probabilités.*

3.3.5 Développement de logiciels

4) **Classification supervisée de courbes :** programme écrit dans le langage S+.
3) **Régression nonparamétrique fonctionnelle :** programme écrit dans le langage S+.
2) **Régression simultanée de courbes :** programme (écrit dans le langage S+) réalisant une ACP de courbes sous contraintes de dimension et de régularité (voir a.3) et disponible à l'adresse *http ://www.lsp.ups-tlse.fr/Fp/Ferraty/SOFTWARES/developpements-de-logiciels.html.*
1) **ACP curvilinéaire :** programme (écrit dans le langage S+) fournissant, de manière interactive, divers graphiques et quantités résultant d'un type d'ACP nonlinéaire (voir a.1 et a.2). Ce programme, appelé *SALSA* (Smoothing Alternating with Least Squares Algorithm), est disponible sur le serveur de la librairie statistique internationale *Statlib*[1] à l'adresse *http ://lib.stat.cmu.edu/S/salsa.*

[1] *Statlib* est une librairie scientifique internationale qui est accessible par le réseau internet. Son objectif est de vulgariser de nouvelles méthodes statistiques au travers de programmes en proposant une utilisation simple et rapide (exemples, démonstrations des possibilités du programme,...).

Chapitre 4

Modélisation Paramétrique

Initialisé conjointement avec Hervé Cardot et Pascal Sarda, nous nous sommes intéressés au modèle fonctionnel de régression linéaire (que l'on peut considérer paramétrique selon la Définition 3.1.3) et qui correspond au modèle de régression linéaire lorsque le régresseur (variable explicative) est une variable aléatoire fonctionnelle (v.a.f.). Nous avons commencé par donner des premières propriétés théoriques d'un estimateur empirique pour ce modèle basé sur l'Analyse en Composantes Principales Fonctionnelle.

Une fois le cadre théorique mis en place, nous avons souligné la nécessité d'utiliser des estimateurs permettant de controler plus finement le lissage. C'est ainsi que nous avons proposé deux nouveaux estimateurs. Le premier consiste à projeter le précédent sur une base B-spline alors que le second fournit un estimateur B-spline directement à partir d'un critère moindres carrés pénalisés. En plus de l'étude asymptotique, nous avons mis en œuvre ces estimateurs et comparé leurs performances respectives selon diverses situations. Enfin, nous avons complété ces travaux en proposant deux statistiques de test portant sur le coefficient fonctionnel du modèle linéaire fonctionnel (en collaboration avec André Mas). La première peut être approximée par une v.a. suivant un χ^2 alors qu'on montre que la loi de la seconde converge vers une gaussienne.

Notons que les outils théoriques mis en jeu dans ce type d'études trouvent leur origine dans l'ACP fonctionnelle (Dauxois et Pousse, 1976, Romain, 1979, Dauxois et *al.*, 1982), la théorie sur les variables aléatoires hilbertiennes (Grenander, 1963, Parthasarathy, 1967), la théorie sur les opérateurs linéaires (Dunford et Schwartz, 1963, Gohberg et Krejn, 1971, Chatelin, 1983, Kato, 1976, Nagy et Riesz, 1965), les inégalités exponentielles pour des variables aléatoires hilbertiennes (Yurinskii, 1976), les notions de distance entre variables aléatoires (Dudley, 1968, Witt, 1974) ainsi que dans les travaux de Bosq (1991) dans le cadre des processus autorégressifs hilbertiens.

4.1 Modèle fonctionnel de régression linéaire (MFRL)

Selon la synoptique du Tableau 3.1, on s'intéresse à la situation (1). Soit X une v.a.f. à valeurs dans l'espace H des fonctions définies sur $\mathcal{C}$ (un compact de $\mathbb{R}$) et de carré intégrable (on parle alors de variables aléatoires hilbertiennes). On a donc $X = \{X(t),\ t \in \mathcal{C}\}$. Soit $(X_i, Y_i)_{i=1,\ldots,n}$ n couples de v.a. identiquement et indépendemmant distribués (i.i.d.) selon le couple (X, Y) où Y une v.a.r. On s'intéresse alors au modèle fonctionnel de régression

$$Y = \int_{\mathcal{C}} \phi(t)X(t)\,dt + \varepsilon, \tag{4.1}$$

où $\phi \in H$ et ε est une v.a.r. centrée de variance σ^2 et telle que $E\left(X(t)\varepsilon\right) = 0$ pour presque tout t. Bien entendu, on peut reformuler de manière équivalente un tel modèle en munissant H du produit scalaire $< .,. >$ défini par $< u, v >= \int_{\mathcal{C}} u(t)v(t)\,dt$:

$$Y = < \phi, X > +\varepsilon \tag{4.2}$$

ou encore

$$Y = \Phi(X) + \varepsilon, \tag{4.3}$$

en supposant que Φ est un opérateur linéaire continu (appartient au dual topologique de H) ; l'équivalenc entre (4.2) et (4.3) est donnée par le théorème de représentation de Riesz. En fin, notons que le modèle fonctionnel de régression linéaire fut introduit historiquement sous la formulation (4.1) (voir Hastie et Mallows, 1993).

4.2 MFRL et Estimations

Dans le cadre de ce modèle, nous avons proposé différents estimateurs. Deux concernent directement l'opérateurs linéaire continu Φ alors qu'un autre s'intéresse uniquement à son représentant dans H à savoir ϕ.

4.2.1 Régression sur composantes principales empiriques

La construction de cet estimateur, qui reprend l'idée de Bosq (1991) dans le cas des processus ARH, utilise le fait qu'il existe une relation entre l'opérateur de variance $\Gamma \;:\; x \in H \;\mapsto\; \Gamma(x) = E(< X, x > X) \in H$ et l'opérateur de covariance $\Delta \;:\; x \in H \;\mapsto\; \Delta(x) = E(< X, x > Y) \in \mathbb{R}$ à savoir

$$\Delta \;=\; \Phi\Gamma.$$

Ainsi, si on dispose d'une "bonne" approximation $\widetilde{\Gamma}$ (resp. $\widetilde{\Delta}$) pour Γ (resp. Δ et si de plus, $\widetilde{\Gamma}$ est inversible, alors on obtient un premier estimateur de Φ^{ACP} de Φ défini par : $\Phi_n^{ACP} = \widetilde{\Delta}\widetilde{\Gamma}^{-1}$. Considérons les versions empiriques de Γ_n de Γ et Δ_n de Δ définies par : $\Gamma_n = 1/n\sum_{i=1}^{n} < X_i, . > X_i$ et $\Delta_n = 1/n\sum_{i=1}^{n} < X_i, . > Y_i$. Soit $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,n} \geq 0 = \lambda_{n,n+1} = \cdots$ les valeurs propres de Γ_n et $V_{n,1}, V_{n,2}, \ldots,$ les fonctions propres associées à ces valeurs propres et posons $\Pi_K = \sum_{k=1}^{K} < V_{n,k}, . > V_{n,k}$ le projecteur sur le sous-espace propre engendré par $\{V_{n,k},\; k = 1\ldots, K\}$. Alors, l'estimateur Φ^{ACP} de Φ est défini par :

$$\Phi_K^{ACP} \;=\; \Delta_n\Pi_K(\Pi_K\Gamma_n\Pi_K)^{-1} \;=\; \sum_{k=1}^{K}\frac{\Delta_n V_{n,k}}{\lambda_{n,k}} < V_{n,k}, . > . \qquad (4.4)$$

On montre alors sous certaines hypothèses ($K = K_n$) que $\sup_{\|x\|=1}|\Phi_K^{ACP}(x) - \Phi(x)|$ converge en probabilité vers 0 quand n tends vers l'infini et aussi en convergence presque sûre en renforçant légèrement les hypothèses ($\|.\|$ désigne la norme associée au produit scalaire $< ., . >$).
D'un point de vue pratique, nous avons étudié le comportement de l'estimateur $\phi_K^{ACP} = \sum_{k=1}^{K}\frac{\Delta_n V_{n,k}}{\lambda_{n,k}}V_{n,k}$ de ϕ en fonction de la taille de l'échantillon choisie dans les simulations. Autant pour une taille importante de l'échantillon, l'estimateur obtenu donne de bons résultats, autant pour de petits échantillon, ϕ_K^{ACP} nécessite une régularisation. C'est l'objet du paragraphe suivant.

Pour plus détails, voir Cardot, Ferraty et Sarda (1999a).

4.2.2 Estimateurs régularisés

Régression sur composantes principales régularisée

L'idée consite à régulariser l'estimateur ϕ_K^{ACP} en utilisant une approximation B-spline (de Boor, 78). Soit $\mathcal{S}_{q,r}$ l'espace des fonctions spline de dimension $q+r$. L'estimateur régularisé $\widehat{\phi}_K^{ACPR}$ est défini par

$$\widehat{\phi}_K^{ACPR} = \arg\min_{\beta \in \mathcal{S}_{q,r}} \int (\phi_K^{ACP}(t) - \beta(t))^2 dt.$$

La convergence en probabilité vers 0 de $\|\widehat{\phi}_{K_n}^{ACPR} - \phi\|_2$ est obtenue quand n tends vers l'infini, où $\|x\|_2 =< \Gamma\, x, x >$. Les vitesses sont données lorsqu'on précise la décroissance des valeurs propres de Γ (géométrique ou exponentielle).

Estimateur développé sur une base de B-splines

Une autre façon d'obtenir un estimateur lisse de ϕ consiste à utiliser un critère des moindres carrés que l'on pénalise en introduisant une contrainte de régularité. Soit $\{B_j\ ;\ j = 1, \ldots, q+r\}$ une base de de B-spline normalisée engendrant $\mathcal{S}_{q,r}$ et $\boldsymbol{B}$ le vecteur formé par les B_j. Alors, on définit l'estimateur en posant $\phi^{SPL}(.) = \sum_{j=1}^{q+r} {}^t\widehat{\theta}_j B_j(.) = {}^t\widehat{\boldsymbol{\theta}}\boldsymbol{B}$ où $\widehat{\boldsymbol{\theta}}$ est solution du problème de minimisation suivant :

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{q+r}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - < {}^t\widehat{\boldsymbol{\theta}}\boldsymbol{B}, X_i > \right)^2 + \rho \left\| {}^t\widehat{\boldsymbol{\theta}}\boldsymbol{B}^{(m)} \right\|^2,$$

ρ étant un paramètre de lissage strictement positif et $\boldsymbol{B}^{(m)}$ le vecteur contenant les dérivées $m^{\text{ème}}$ de chaque composante de $\boldsymbol{B}^{(m)}$. La vitesse de convergence en moyenne quadratique de $\|\phi^{SPL} - \phi\|_2$ conditionnellement à X_1, X_2, ..., X_n est donnée.

Commentaires

D'un point de vue pratique, le choix optimal des paramètres (dimension, lissage) est obtenu en utilisant le critère de validation croisée généralisée (voir Marx et Eilers 1999 qui propose l'usage de ce critère dans un contexte similaire). Lors de simulations, nous n'avons pas constaté de comportement systématiquement meilleur d'un estimateur par rapport à l'autre bien que d'un point de vue théorique, ϕ^{SPL} converge plus rapidement. Enfin, notons que nous avons appliqué le MFRL sur les données agronomiques présentées au Chapitre 1.

L'ÉTUDE APPLIQUÉE SE TROUVE DANS CARDOT, FERRATY ET SARDA (1999B) ; LES DÉTAILS D'ORDRE THÉORIQUE SE TROUVENT DANS CARDOT, FERRATY ET SARDA (2003).

4.3 MFRL et Tests

Dans la poursuite naturelle des travaux exposés précédemment et ceci avec en plus la collaboration d' André Mas, nous avons mis en place deux procédures de test permettant de tester l'hypothèse nulle $\phi = 0$ contre son alternative $\phi \neq 0$, chacune basée sur la décomposition spectrale de l'opérateur Γ_n. On définit alors deux statistiques de test, notées D_n et T_n, par :

$$D_n = \frac{1}{\widehat{\sigma}_n^2} \left\| \sqrt{n}\Delta_n G_n \right\|^2 \quad \text{et} \quad T_n = \frac{1}{\sqrt{K_n}} \left(\frac{1}{\widehat{\sigma}_n^2} \left\| \sqrt{n}\Delta_n G_n \right\|^2 - K_n \right),$$

avec $G_n = \sum_{k=1}^{K_n} \lambda_{n,k}^{-1/2} < V_{n,k}, . > V_{n,k}$ et $\widehat{\sigma}_n^2$ un estimateur de σ^2 (dans la pratique, on prendra $\widehat{\sigma}_n^2 = 1/tr(\boldsymbol{I}_n - \boldsymbol{S}) \sum_{i=1}^{n} \left(Y_i - < \phi^{SPL}, X_i >\right)^2$ où $\boldsymbol{S}$ est la "hat matrix"). Nous avons montré que T_n converge en distribution vers une loi gaussienne $N(0, 2)$ alors que D_n peut être approximée pour n suffisamment grand, par un χ^2 ayant k_n degrés de liberté. Par ailleurs, ces deux statistiques de tests on donné des résultats similaires et satisfaisants lors de simulations portant sur le niveau et la puissance de tels tests.

POUR PLUS DE PRÉCISIONS, VOIR CARDOT, FERRATY, MAS ET SARDA (2003)

Chapitre 5

Modélisations NonParamétriques

Parallèlement aux développements paramétriques, je me suis intéressé en collaboration avec Philippe Vieu aux modèles fonctionnels nonparamétriques. Ce domaine de recherche est d'autant plus motivant qu'aucune étude n'avait jusqu'alors abordé cette problématique. Il s'agit alors de défricher mais aussi de proposer de nouveaux développements, à la fois théoriques et pratiques. Á la différence de ce qui se passe dans le cadre fonctionnel paramétrique, l'objet que l'on souhaite estimer n'est pas indexable par un nombre fini de paramètres fonctionnels. Ceci nous conduit à proposer des extensions des méthodes nonparamétriques multivariées capable de prendre en compte de v.a.f.

Différentes problématiques ont été abordées. En particulier, le modèle fonctionnel nonparamétrique de régression. Il tient une place prépondérante car nos premières investigations ce sont portées sur ce type de modélisation. Le cadre théorique, les premiers résultats asymptotiques ainsi que les premières applications ont été fournis dans ce contexte. Par la suite, nous avons réalisé des extensions permettant d'englober l'étude de séries temporelles. D'autres problématiques ont motivé nos travaux. En effet, nous avons développé une méthode nonparamétrique de classification supervisée (discrimination) de courbes dont la consistance se déduit des études concernant le modèle fonctionnel nonparamétrique de régression et qui donne satisfaction d'un point de vue pratique. Enfin, nous avons abordé un autre problème statistique de nature nonparamétrique qui concerne l'étude de la distribution d'une v.a.r. conditionnellement à une v.a.f.

Du point de vue des outils mis en jeu, on est amené à utiliser des inégalités exponentielles, que ce soit pour des sommes de v.a.r. indépendantes (Uspensky, 1937) ou dépendantes (Rio, 1999), mais aussi des notions de dimension fractale pour des processus (Pesin, 1993) ainsi que tout l'arsenal dont on dispose sur les estimateurs à noyaux (Schimek, 2000).

5.1 Modèles NonParamétriques de Régression

5.1.1 Modèle NonParamétrique de Régression : cas indépendant

Commençons par présenter le modèle fonctionnel nonparamétrique de régression. Soit X une v.a.f. à valeurs dans un espace semi-métrique (E, d) ,Y une v.a.r. et $(X_i, Y_i)_{i=1,\ldots,n}$ n couples identiquement et indépendamment distribués suivant (X, Y). On s'intéresse alors au modèle

$$Y_i \;=\; r(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

où les ε_i sont des v.a.r. centrées telles que $E(\varepsilon_i/X_i) = 0$. Ce qui change fondamentalement par rapport au modèle fonctionnel paramétrique de régression, c'est qu'ici r doit vérifier la condition de type Lipschitz suivante

$$\exists C > 0, \exists\beta > 0, \forall(x, y) \in E \times E, |r(x) - r(y)| \leq Cd(x, y)^\beta \tag{5.2}$$

au lieu de supposer r opérateur linéaire continu. Notons que la condition (5.2) est satisfaite pour $\beta = 1$ dès que r est linéaire continu. L'objectif est de définir un estimateur consistant de r. En s'inspirant de ce qui se fait dans le cadre vectorielle, on propose un estimateur à noyau $\widehat{r}$ de r défini par :

$$\widehat{r}(x) \;=\; \frac{\sum_{i=1}^n Y_i\, K\left(d(x, X_i)/h\right)}{\sum_{i=1}^n K\left(d(x, X_i)/h\right)},$$

où K est un noyau et h la largeur de fenêtre. Contrairement au cas vectoriel, il est difficile de supposer que X admettent une densité eventuellement strictement positive au point d'estimation (sauf pour des v.a.f. X particulières : sous certaines conditions, les processus de diffusion admette une densité par rapport à la mesure de Wiener, voir Lipster and Shiryayev, 1972). L'idée consiste alors à obtenir des résultats asymptotiques en fonction de la quantité $P\left(d(x, X) < h\right)$; selon le comportement de cette quantité en fonction de

h, on en déduira différentes vitesses de convergence. De façon générale, on suppose

$$\exists(\beta_1, \beta_2) \in \mathbb{R}^{+2},\ 0 < \beta_1 \phi_x(h) \leq P\left(X \in \mathcal{B}(x, h)\right) \leq \beta_2 \phi_x(h). \tag{5.3}$$

Par exemple, en rajoutant d'autres hypothèses classiques en estimation non-paramétriques, on obtient le résultat ponctuel suivant :

$$\widehat{r}(x) - r(x) \ = \ O(h^\beta) + O\left(\sqrt{\frac{\log n}{n\, \phi_x(h)}}\right), p.s.$$

Le terme de biais $O(h^\beta)$ provient essentiellement de la régularité de r alors que le second $O\left(\sqrt{\log n/(n\, \phi_x(h))}\right)$ dépend de la concentration de la mesure de probabilité sur des petites boules (voir Li and Shao 2001 pour des calculs de probabilités obtenus sur des processus standards). En particulier, si $P\left(d(x, X) < h\right) \sim C_x\, h^{\delta(x)}$, ce qui signifie que la distribution de X admet une dimension fractale ponctuelle égale à $\delta(x)$ (voir Pesin, 1993), on peut préciser la vitesse de convergence. Bien entendu, en renforçant les hypothèses, on est capable de donner des résultats uniformes.

Enfin, d'un point de vue pratique, cet estimateur a donné de bon résultats concernant l'estimation du taux de matière contenu dans des morceaux de viandes à partir de courbes spectrométriques.

Pour plus de précisions, voir Ferraty and Vieu (2000), Ferraty et Vieu (2002)

5.1.2 Extension aux séries temporelles

Considérons $Z = \{Z(t)\ ;\ t \in \mathcal{T}_N\}$ un processus réel stationnaire à temps continu observé sur le compact $\mathcal{T}_N = [0;\ N\tau[$. À partir de Z, on construit N v.a.f. $(W_i)_{i=1,\dots,n}$ de la façon suivante :

$$\forall t \in [0;\ \tau[,\ W_i(t) = Z\left((i-1)\tau + t\right),\ i = 1, \dots, N.$$

Soit $G(W_{i+1})$ une caractéristique future de W_i. Alors, on peut modéliser le processus Z à l'aide du modèle fonctionnel suivant

$$G(W_{i+1}) \ = \ \Psi(W_i) + \varepsilon_i,\ i = 1, \dots, N, \tag{5.4}$$

où G est connu, Ψ est inconnu (à estimer) supposé satisfaire la condition de Lipschitz (5.2) et les ε_i's sont des v.a.r. centrées. De nombreuses applications peuvent se déduire de ce type de modélisation. En effet, on peut prendre $G(W_{i+1}) = W_{i+1}(T)$ où $T \in [0;\ \tau[$; on s'intéresse alors à la prédiction d'une valeur future (ce qui est le cas pour les données de consommation d'électricité). Si on pose $G(W_{i+1}) = \sup_{t\in[0;\ \tau[} W_{i+1}(t)$, on se focalise sur la prévision du supremum (par exemple, prévision de pics de pollution d'ozone, Aneiros Perez et *al.*, 2002). Mais on peut aussi prendre $G(W_{i+1}) = (W_{i+1}(\tau) - W_{i+1}(0))/W_{i+1}(0)$, ce qui correspond à la prévision d'un taux de variation sur un intervalle, ce qui peut être utile lorsqu'on étudie la tendance sur une courte durée d'un indice économique.

Revenons maintenant à l'aspect modélisation et posons $X_i = W_i$ et $Y_i = G(W_{i+1})$; il est clair que (5.4) est un modèle fonctionnel de régression. À la différence avec le modèle (5.1), c'est la dépendance évidente qu'il existe entre les couples (X_i, Y_i) que l'on modélise en supposant que la séquence $\{(X_i, Y_i)\}_i$ est α-mélangeante. Enfin, en s'inspirant de ce qui a été fait dans le cas indépendant, on propose d'étudier l'estimateur à noyau $\widehat{\Psi}$ de Ψ suivant :

$$\widehat{\Psi}(x) \;=\; \frac{\sum_{i=1}^{n} G(W_{i+1})\, K\left(d(x, W_i)/h\right)}{\sum_{i=1}^{n} K\left(d(x, W_i)/h\right)}.$$

Ici aussi, la clé de l'obtention de résultats de convergence réside, en plus de l'hypothèse de type (5.3) sur la distribution de W dont sont issus les W_i, dans un autre hypothèse de concentration portant sur la distribution conjointe des (W_i, W_j) :

$$\exists(\eta_1, \eta_2) \in \mathbb{R}^{+2},\ 0 < \eta_1 \zeta_x(h) \le \sup_{i\neq j} P\left(\{d(x, W_i) < h\} \cap \{d(x, W_j) < h\}\right) \le \eta_2 \zeta_x(h).$$

On obtient alors des résultats asymptotiques du type

$$\widehat{\Psi}(x) - \Psi(x) \;=\; O(h^{\beta}) + O\left(\sqrt{\frac{\log n}{n\, \phi_x(h)}}\right) + O(U_n),\ p.s.,$$

où U_n dépend de $\phi_x(h)$, de $\zeta_x(h)$ et de la structure de mélange.

Pour plus de détails, voir Ferraty, Goia et Vieu (2002a et 2002c)

5.1.3 Probabilités de petites boules et vitesses de convergence

Nous avons déjà souligné que les vitesses de convergence dépendent de la quantité $P(d(x,X) < h)$. Ces vitesses peuvent être alors précisées en utilisant les résultats sur les probabilités de petites boules pour des processus standards. De façon générale, pour des processus de diffusion (admettant une densité par rapport à la mesure de Wiener) et gaussiens, on a des approximations (pour plus de détails, voir Li and Shao, 2001) du type $P(d(x,X) < h) \sim C_x\, e^{-C_1 h^{-C_2}}$ où C_x, C_1 et C_2 sont des constantes positives. Ceci nous permet de déduire pour nos estimateurs des vitesses en $O\left((\log n)^{-C_3}\right)$ où C_3 est une constante positive qui dépend de C_2 mais aussi de la métrique d choisie. Ce résultat n'est pas surprenant dans un tel contexte doublement fonctionnel. Cependant, d'un point de vue statistique, on peut s'intéresser à des semi-métriques d qui augmentent la concentration de X autour de x, ce qui s'obtient aisément, par exemple, en projetant sur un sous-espace de dimension finie k. On a alors que $P(d(x,X) < h) > C_x\, h^k$ et on en déduit pour nos estimateurs une vitesse de convergence de l'ordre de

$$(n/\log n)^{-2C_4/(2C_4+k)} \quad p.co.$$

Pour plus de détails, voir Ferraty, Laksaci et Vieu (2003)

5.1.4 Travaux de synthèses

Les études précédentes nous ont amené à rédiger des ouvrages de synthèses portant sur les modèles régression nonparamétriques :

Ferraty et Vieu (2001), Ferraty, Núñez Antón and Vieu (2001), Ferraty, Goia and Vieu (2002b).

5.2 Discriminations de courbes

Les modèles précédemment présentés s'inscrivent dans le cadre général des modèles d'espérance conditionnelle ($E(Y/X) = r(X)$) où Y est une v.a.r. Un prolongement naturel est de voir ce qui se passe lorsque la variable Y est qualitative à valeurs dans $\overline{C} = \{1,\ldots,C\}$ définissant ainsi C classes ou groupes.

On se situe alors dans la problématique de la classification supervisée (ou discrimination) de courbes (lorsque l'observation de la v.a.f. X est une courbe) et on s'intéresse à l'estimation des probabilités $p_c(x) = P(Y = c/X = x)$ pour $c \in \overline{C}$. En remarquant que

$$E\left(1_{[Y=c]/X=x}\right) = P(Y = c/X = x),\ c \in \overline{C},$$

où $1_{[Y=c]} = 1$ si $Y = c$ et 0 sinon, on déduit aisément un estimateur à noyau $\widehat{p}_{c,h}(x)$ de $p_c(x)$ défini par :

$$\widehat{p}_{c,h}(x) = \frac{\sum_{i=1}^n 1_{[Y_i=c]} K\left(d(x, X_i)/h\right)}{\sum_{i=1}^n K\left(d(x, X_i)/h\right)}.$$

Cet estimateur hérite des propriétés asymptotiques de l'estimateur de la régression. En effet, en supposant que les fonctions $p_c(.)$ vérifient la condition

$$\exists C_c, \exists \beta > 0, \forall (x, y) \in E \times E,\ |p_c(x) - p_c(y)| \leq C_c d(x, y)^\beta,$$

et sous l'hypothèse de concentration (5.3), en plus des hypothèses usuelle protant sur K et h, on obtient le résultat suivant

$$\widehat{p}_{c,h}(x) - p_c(x) = O\left(h^\beta\right) + O\left(\sqrt{\frac{\log n}{n\,\phi_x(h)}}\right), \quad p.s.$$

D'un point de vue pratique, on choisit la semi-métrique dans une famille indexée par un paramètre ; une procédure de type validation croisée nous permet de sélectionner une semi-métrique ainsi qu'une fenêtre h. Les performances de cet estimateur sont illustrées notamment dans le cadre de la reconnaissance vocale.

Pour plus de détails, voir Ferraty et Vieu (2003a)

5.3 Distribution et conditionnement par une v.a.f.

Une autre façon d'étudier le lien entre une v.a.r. Y et une v.a.f. X est de s'intéresser à la fonction de répartition de Y sachant X définie par :

$$F^x(y) = P\left(Y < y | X = x\right).$$

En s'inspirant des idées précédemment développées dans le cadre fonctionnel ainsi que des études réalisées dans le cas vectoriel (voir Roussas, 1969 et Samanta, 1989), un estimateur $\widehat{F^x}(y)$ de $F^x(y)$ peut être défini de la manière suivante :

$$\widehat{F^x}(y) = \frac{\sum_{i=1}^n K\left(d(x,X_i)/h\right) H\left((y-Y_i)/h_H\right)}{\sum_{i=1}^n K\left(d(x,X_i)/h\right)}, \tag{5.5}$$

où k est une largeur de fenêtre associée au noyau K, H une fonction de répartition connue et h_H un paramètre strictement positif. Si on suppose que F^x vérifie la condition

$$\forall (y_1,y_2) \in \mathbb{R}\times\mathbb{R}, \forall (x_1,x_2) \in E\times E, |F^{x_1}(y_1) - F^{x_2}(y_2)| \leq C_x \left(d(x_1,x_2)^b + |y_1-y_2|^b\right)$$

et que X satisfasse l'hypothèse de concentration (5.3), alors on obtient le résultat suivant :

$$\widehat{F^x}(y) - F^x(y) \;=\; O(h^b) + O(h_H^b) + O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right). \qquad p.s.$$

Plusieurs applications découlent de cet estimateur. D'une part, on peut obtenir une estimation des quantiles de Y conditionnellement à X. D'autre part, on peut déduire de $\widehat{F^x}(y)$ un estimateur de la densité conditionnelle $f^x(y)$. En effet, on définit cet estimateur noté $\widehat{f^x}(y)$ en dérivant (5.5) :

$$\widehat{f^x}(y) = \frac{h^{-1}\sum_{i=1}^n K\left(d(x,X_i)/h\right) H'\left((y-Y_i)/h_H\right)}{\sum_{i=1}^n K\left(d(x,X_i)/h\right)}.$$

Il suffit alors de poser une condition de régularité portant sur $f^x(y)$ du type

$$\forall (y_1,y_2) \in \mathbb{R}\times\mathbb{R}, \forall (x_1,x_2) \in E\times E, |f^{x_1}(y_1) - f^{x_2}(y_2)| \leq C_x \left(d(x_1,x_2)^{b_1} + |y_1-y_2|^{b_2}\right)$$

ainsi que l'hypothèse de concentration (5.3) pour obtenir un résultat du type

$$|\hat{f}^x(y) - f^x(y)| \;=\; O(h^{b_1}) + O(h_H^{b_2}) + O\left(\sqrt{\frac{\log n}{n\, h_H\, \phi_x(h)}}\right), \qquad p.s.$$

Pour plus de détails, voir Ferraty, Laksaci et Vieu, 2003

5.4 Modèle à indice fonctionnel simple ; un premier pas vers les modèles fonctionnels semiparamétriques

On s'intéresse à un modèle fonctionnel de régression d'une v.a.r. Y sur une v.a.f. X à valeur dans $(E, < .,. >)$. C'est donc une extension naturelle des modèles à indice simple (Delecroix et Hristache, 1999, Härdle et *al.*, 1993) au cadre fonctionnel. Ce modèle se présente sous la forme suivante :

$$Y \; = \; r(< X, \theta_0 >) + \varepsilon,$$

où r est une fonction réelle inconnue, θ_0 un paramètre fonctionnel (car élément de E) inconnu et ε une v.a.r. centrée telle $E(\varepsilon|X) = 0$. Nous avons d'abord étudié ce modèle pour θ_0 fixé. On s'intéresse alors à un estimateur de $r_{\theta_0}(.) = r(< ., \theta_0 >)$, noté $\widehat{r}_{\theta_0}$ et défini par

$$\widehat{r}_{\theta_0}(x) \; = \; \frac{\sum_{i=1}^n Y_i K\left(< X_i - x, \theta_0 > /h\right)}{\sum_{i=1}^n K\left(< X_i - x, \theta_0 > /h\right)}.$$

En supposant des hypothèses classiques, on obtient un résultat du type

$$\widehat{r}_{\theta_0}(x) - r_{\theta_0}(x) \; = \; O\left(h^\beta\right) + O\left(\sqrt{\frac{\log n}{n\, P(| < X - x, \theta_0 > | < h}}\right).$$

Bien entendu, il reste à faire l'étude lorsque θ_0 n'est pas fixé. Une autre façon de voir ce modèle est de dire qu'il s'agit d'un modèle fonctionnel de régression nonparamétrique pour lequel $d_{\theta_0}(x, y) = | < x - y, \theta_0 > |$. Autrement dit, c'est une manière de déterminer une semi-métrique optimale parmi une famille de semi-métriques indexée par un paramètre fonctionnel.

Pour plus de détails, voir Ferraty, Peuch et Vieu (2003).

Chapitre 6

Quelques applications

6.1 Courbes de pluviométrie et de températures : étude d'une production de blé

L'objectif de cette étude est de mettre en relation le rendement de blé avec les indicateurs météorologiques tels que la pluviométrie et la température. Dans ce but, on dispose de 198 récoltes de blé qui fournissent 198 rendements notés $Y_1, \ldots, Y_198$. On dispose des températures cumulées journalières (T) et des précipitations cumulées journalières (P) mesurées du 1er Octobre jusqu'à fin Août (voir Figure 3.2). Enfin, pour chaque production i, la durée de culture L_i a été observée. Afin d'expliquer le rendement de blé, on a considéré le modèle fonctionnel de régression linéaire suivant :

$$Y_i \ = \ aL_i + \int \alpha_1(t) T_i(t)\, dt + \int \alpha_2(t) P_i(t)\, dt + \varepsilon_i, \quad i = 1, \ldots, 198.$$

On doit donc estimer le paramètre réel a associé à la durée de culture L, le paramètre fonctionnel α_1 associé aux courbes de températures et enfin la pramètre fonctionnel α_2 associé aux courbes de précipitations en résolvant le problème suivant de minimisation :

$$\min_{a,\alpha_1,\alpha_2} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - aL_i - \int \alpha_1(t) T_i(t)\, dt - \int \alpha_2(t) P_i(t)\, dt \right)^2 + \rho_1 \|\alpha_1^{(2)}\|^2 + \rho_2 \|\alpha_2^{(2)}\|^2,$$

où $(\alpha_1, \alpha_2) \in S_{q,k} \times S_{q,k}$ avec $q = 4$ et $k = 25$. Les choix optimaux de ρ_1 et ρ_2 sont obtenus à l'aide de la validation croisée (voir Figure 6.2). Concernant les paramètres fonctionnels (α_1 et α_2), on obtient les résultats suivants :

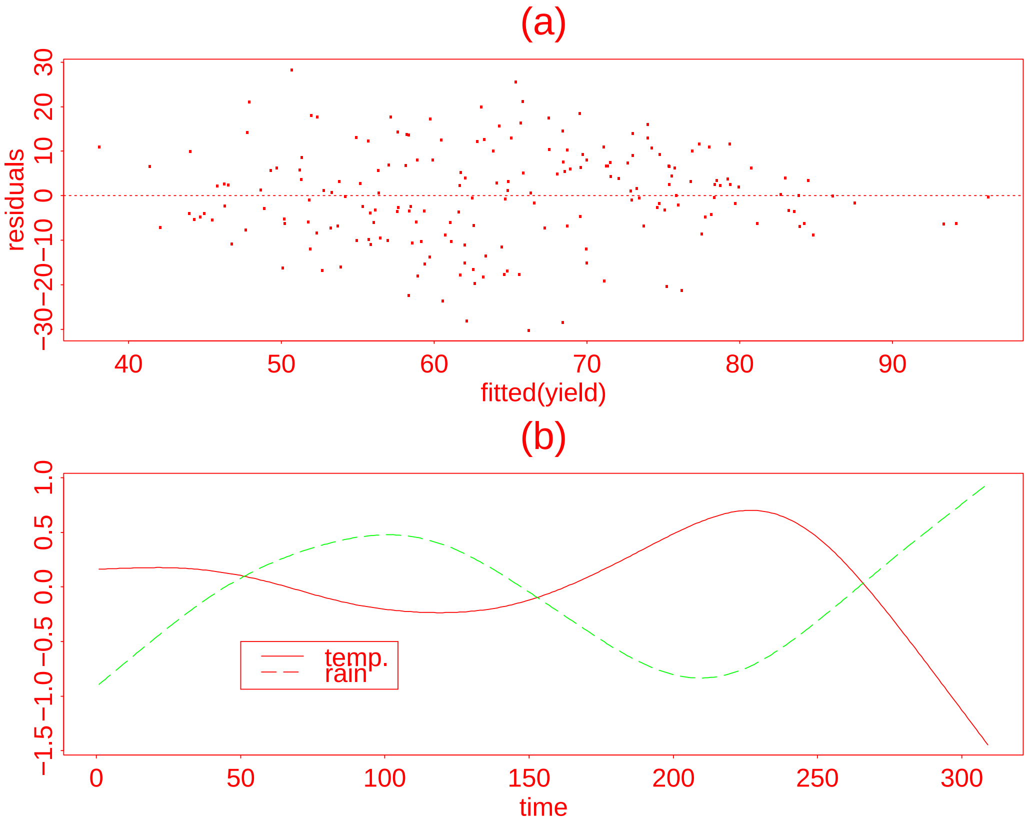

FIG. 6.1 – Validation croisée pour choisir ρ_1 et ρ_2 ; le minimum est atteint pour $\rho_1 \simeq \rho_2 \simeq 0.001$

On remarque en particulier sur la Figure 6.1(b) que des températures élevées et/ou peu de précipitations durant le printemps et le début de l'été contribuent à augmenter le rendement.

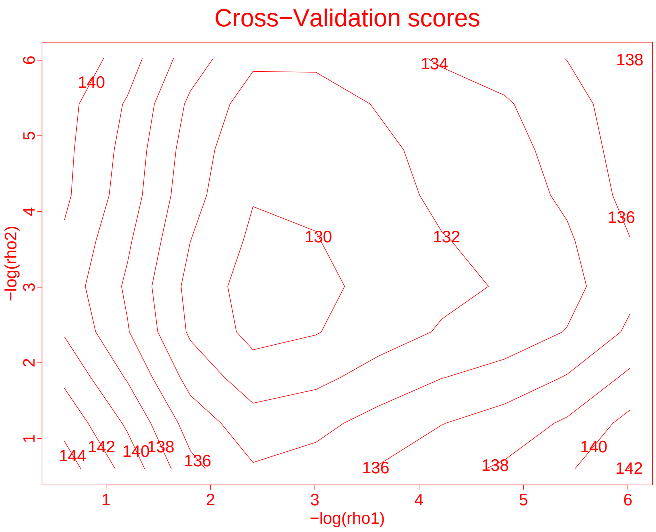

FIG. 6.2 – Validation croisée pour choisir ρ_1 et ρ_2 ; le minimum est atteint pour $\rho_1 \simeq \rho_2 \simeq 0.001$

6.2 Courbes spectrométriques : estimation du taux de matières grasses

On reprend ici les données représentées dans la Figure 3.1. On se place dans le cadre du modèle fonctionnel de régression nonparamétrique $Y_i = r(X_i) + \varepsilon_i$, pour $i = 1, \ldots, n$ où Y_i est le taux de lipide obtenu après une analyse chimique et X_i la courbe spectrométrique correspondante. L'estimateur de r est donné par

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(d_m(x, X_i)/h(x)\right)}{\sum_{i=1}^{n} K\left(d_m(x, X_i)/h(x)\right)},$$

où $d_m(x, y) = \sqrt{\int \left(x^{(m)} - y^{(m)}(t)\right)^2}$. Le choix de cette famille de semi-métriques a été guidé par des avis de chimiométriciens mais aussi par le graphique suivant :

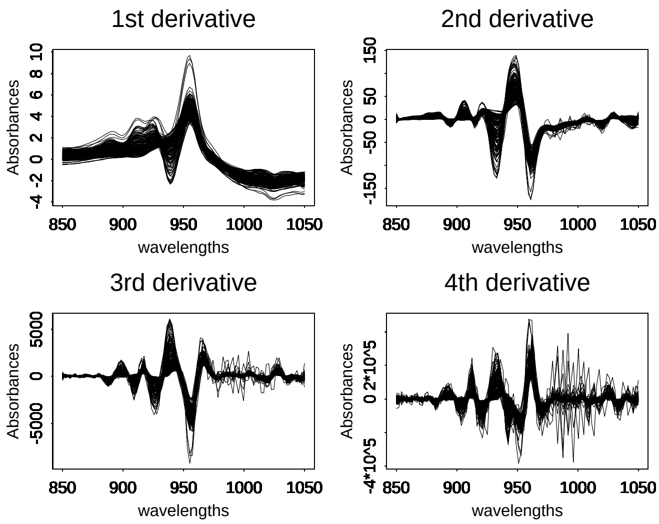

FIG. 6.3 – Derivées successives des courbes spectrométrique

Le choix de la fenêtre locale $h(x)$ ainsi que celui du paramètre m se fait par validation croisée ; on a obtenu $m_{opt} = 2$. Par ailleurs, pour un échantillon test de 50 courbes, on a les estimations suivantes :

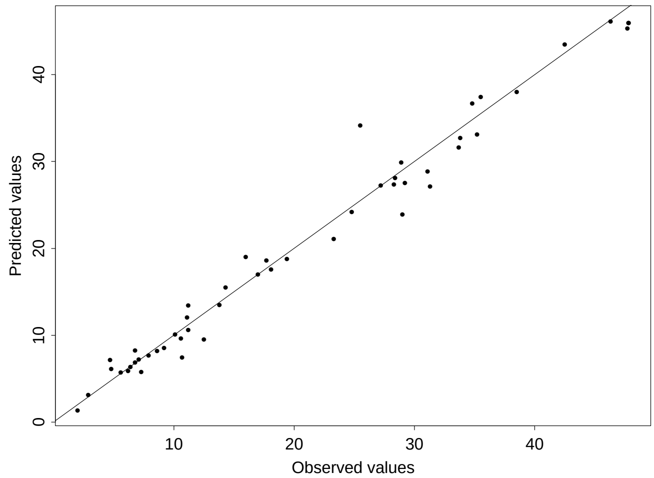

FIG. 6.4 – Comparaison entre les valeurs observées et les valeurs estimées

6.3 Prévisions de consommations électriques

On s'intéresse à des données mensuelles de consommation électrique aux Etats-Unis de Janvier 1973 à Février 2001 (338 mois). La Figure 6.5 fournit la série stationnarisée (la différentiation du log des données initiales permet déliminer l'hétéroscédasticité ainsi que la tendance linéaire).

Cet exemple permet d'illustrer le modèle fonctionnel de régression nonparamétrique appliqué aux séries temporelles. En pratique, on découpe notre série en 28 morceaux correspondant à 28 années ; les 26 premiers constituent l'échantillon d'apprentissage. En reprenant les notations du Paragraphe 5.1.2, on a $N = 26$, $\tau = 12$ et on prédit les 12 valeurs futures en prenant successivement pour $G(W_{27})$ les 12 quantités $\{W_{27}(j)\}_{j=1,\dots,12}$. Les consommation électriques des 12 mois suivants $\{G(W_{28}) = W_{28}(j)\}_{j=1,\dots,12}$ sont prédites de la même façon en incluant les 12 valeurs prédites précédemment. Quant à la semi-métrique, on a choisit la famille suivante

$$d_m(x_1, x_2) \;=\; \left(\int (\Pi_m[x_1(t) - x_2(t)])^2 \; dt\right)^{1/2},$$

où Π_m est le projecteur orthogonal sur le sous-espace propre de dimension m de l'opérateur de covariance empirique des W_i's. Le choix de m se fait par validation croisée et Figure 6.6 compare les 24 mois prédits avec leur valuers observées.

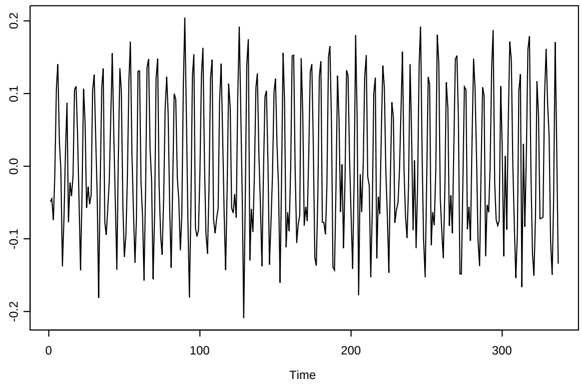

FIG. 6.5 – Données de consommation électrique

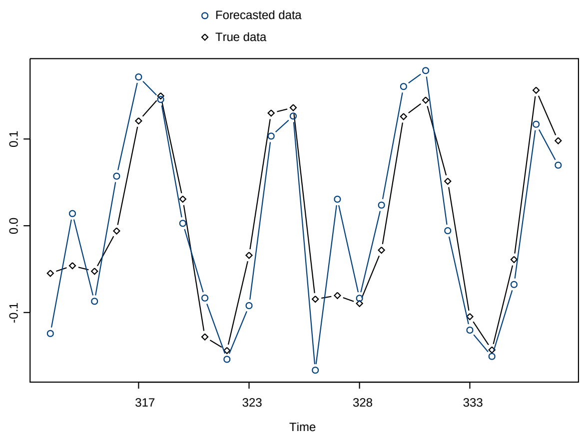

FIG. 6.6 – Données de consommation électrique prédites

6.4 Courbes phonétiques : classification supervisée de phonèmes

On considère les courbes phonétiques représentées Figure 3.5. Ces données nous permettent d'illustrer la méthode nonparamétrique de discrimination de courbes présentée au Paragraphe 5.2. Dans ce but, on considère un echantillon d'apprentissage de courbes phonétiques contenant 750 log-périodogrammes, 150 par classe. À partir d'autres log-périodogrammes, on construit 50 échantillons test, chacun étant composé de 1250 courbes phonétiques, 250 par classe. On a obtenu alors les taux de mauvais classements suivants :

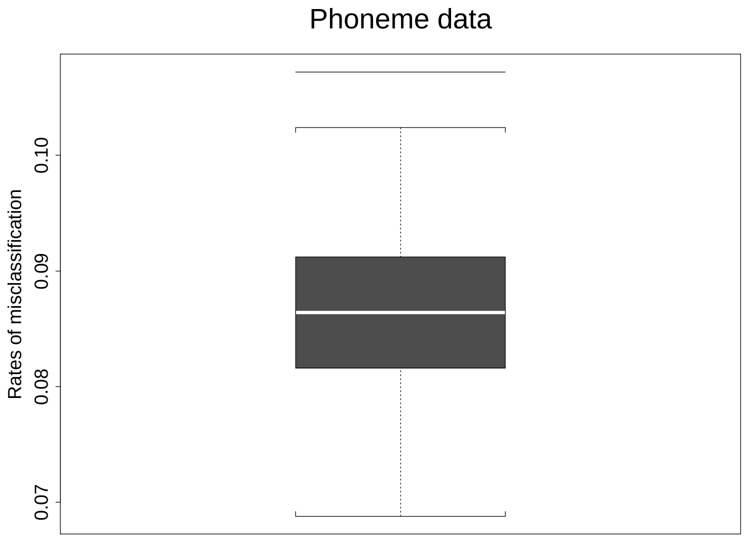

FIG. 6.7 – Taux de mauvais classement

Ajoutons ici que la famille de semi-métriques qui donne le meilleur résultat est basé sur la méthode “multivariate partial least squares regression” (Martens et Naes, 1989).

Chapitre 7

Études en cours, problèmes ouverts et conclusion

La richesse de ce domaine de recherche offre de nombreuse perspectives, tant d'un point de vue théorique que pratique. Dans ce qui suit, je propose une liste de sujets dont certains sont en cours de développement.

7.1 Études en cours

1) **Courbes modales :**
Ce travail en collaboration avec S. Niang et P. Vieu porte sur l'estimation de courbes modales. Considérant un échantillon $X_1,...,X_n$ d'une v.a.f. X, on propose d'estimer sa densité par l'estimateur $\widehat{f}$ suivant :

$$\widehat{f}_h(x) \; = \; \frac{1}{n\,h^\delta}\sum_{i=1}^{n} K\left(d(x,X_i)/h\right),$$

où le paramètre δ provient de l'hypothèse de concentration portant sur X ($\lim_{h\to 0} P(d(x,X) < h)/h^\delta = C_x$). Un premier estimateur empirique X_{mod} de la courbe modale consiste à choisir parmi l'échantillon, la courbe qui maximise la densité :

$$X_{mod} \; = \; \arg\min_{x\in\{X_1,\ldots,X_n\}} \widehat{f}_h(x).$$

L'existence de densité de v.a.f. n'est pas systématique. Cependant on sait, par exemple, que certains processus de diffusion admettent une

densité par rapport à la mesure de Wiener (pour plus de détails, voir Niang 2002b, chapitre 1).
Notons que la détermination de la courbe modale ne nécessite pas la connaissance de la dimension fractale δ. Quant au choix de la fenêtre h, il est basé sur un critère de stabilité. Ajoutons que la détermination de courbes modales peut aussi avoir des applications en terme de classification non supervisée de courbes.

2) **Processus spatio-temporel :**
Une première étude en collaboration avec L. Cornez et P. Sarda concerne la dynamique et la modélisation de l'occupation du sol appliquées à un anthroposystème montagnard des Pyrénées ; il s'agit d'un projet en collaboration avec le laboratoire GEODE (Géographie de l'environnement) de l'Université Toulouse II. L'objectif est de prévoir pour chaque pixel d'une carte le type de végétation v ($\in \{1, \ldots, V\}$), cette zone géographique étant observée à différentes. On modélise alors la probabilité de transition d'un pixel d'un état à la date t (noté Z^t) à un autre état à la date $t+1$ à l'aide d'un modèle polychotomique :

$$P\left(Z^t = v | D_Z^{t-1}, Y^1, \ldots, Y^J\right) = \frac{\exp \theta(v | D_Z^{t-1}, Y^1, \ldots, Y^J)}{\displaystyle\sum_{v'=1}^{V} \exp \theta(v' | D_Z^{t-1}, Y^1, \ldots, Y^J)},$$

avec

$$\begin{aligned} \theta(v | D_Z^{t-1}, Y^1, \ldots, Y^J) &= \theta(v | D_Z^{t-1}, Y^1, \ldots, Y^J, \delta) \\ &= \alpha_v + \textstyle\sum_{z \in D_Z^{t-1}} \sum_{l=1}^{V} \beta_{vl} 1_{[z=l]} + \sum_{j=1}^{J} \gamma_{vj} Y^j, \end{aligned}$$

où D_Z^t est un voisinage autour de Z^t, $Y^1, \ldots, Y^J$ un ensemble de variables topographiques et δ le vecteur contenant les paramètres inconnus de ce modèle :

$$\begin{aligned} \delta = \ & (\alpha_1, \ldots, \alpha_{V-1}, \beta_{11}, \ldots, \beta_{1V}, \beta_{21}, \ldots, \beta_{2,V}, \ldots, \beta_{V-1,1}, \ldots, \beta_{V-1,V}, \\ & \gamma_{11}, \ldots, \gamma_{1J}, \ldots, \gamma_{V-1,1}, \ldots, \gamma_{V-1,J}) . \end{aligned}$$

L'estimation des paramètres se fait en maximisant une vraisemblance pénalisée, ce qui se justifie en reprenant les mêmes arguments que Kooperber, Bose et Stone (1997).

3) **Modèle à indice fonctionnel simple :**
Comme cela a été suggéré au Paragraphe 5.4, l'intérêt d'un tel modèle

consiste en l'estimation non seulement de r mais aussi de l'indice fonctionnel θ_0. Nous envisageons de réaliser cela en utilisant un critère de validation croisée du type $\widehat{\theta}_0 = \arg\min_\theta \sum_{i=1}^n \left(Y_i - \widehat{r_\theta}^{-i}\right)^2$ et en restreignant θ à un espace de fonctions splines par exemple.

4) **Modèle fonctionnel de régression nonparamétrique et normalité asymptotique :**
Ce travail en collaboration avec A. Mas et P. Vieu consiste à écrire des expressions précises des termes de biais et de variance (notamment les constantes) afin d'obtenir la normalité asymptotique de notre estimateur. Ainsi, cela nous permettrait la mise en place de tests basés justement sur ces distributions asymptotiques.

7.2 Autres problèmes ouverts

De nombreuses discussions au sein de STAPH ont permis de dégager les différentes problématiques suivantes :

1) **Modèle fonctionnel paramétrique ou nonparamétrique ?**
   1.1) Comparaison des résultats asymptotiques : l'idée consiste à obtenir des résultats sur les probabilités de petites boules en fonction de la décomposition spectrale de l'opérateur de covariance.
   1.2) Tester paramétrique contre nonparamétrique.
2) **Modèle fonctionnel de régression linéaire dans le cas dépendant :**
Il s'agit d'étendre les résultats obtenus dans le modèle fonctionnel de régression linéaire au cas où les erreurs sont supposées dépendantes.
3) **Modèle fonctionnel de régression linéaire locale :**
Nombreux sont ceux qui se sont intéressés aux modèles de régression basés sur la méthode des polynôme locaux (voir Fan et Gijbels, 1996). Une extension assez naturelle consiste à considérer un régresseur de type v.a.f. ; on s'intéresse alors au problème suivant

$$\min_{\theta_0(x),\theta_1(x)} \sum_{i=1}^n \left\{ (Y_i - \theta_0(x) - < \theta_1(x), X_i >)^2 K\left(d(x, X_i)/h\right) \right\}.$$

L'estimation de la fonction de régression r au point x est donnée par $\widehat{r}(x) = \widehat{\theta}_0(x) + < \widehat{\theta}_1(x), x >$.
4) **Modèles partiellement linéaires :**
Les économétriciens se sont largement intéressés à ce type de modèles

(voir dans un cadre classique Aneiros Pérez, 2001 pour la bibliographie et pour des développements récents, Golubev et Härdle, 2002). On pourrait alors s'intéresser à une version fonctionnelle de la forme

$$Y_i \ = < \theta, Z_i > +m(X_i) + \varepsilon_i,$$

où Y est une v.a.r., X et Z 2 v.a.f., θ et m étant à estimer.

5) **Processus spatio-temporel fonctionnel :**
L'objectif à moyen terme consiste à généraliser l'étude concernant les processus spatio-temporel pour des images satellites observées à des instants rapprochés, ce qui nous amène à un cadre fonctionnel.

7.3 Conclusion

Pour conclure ce document de synthèse, je me contenterai de rappeler deux citations qui recadrent le contexte dans lequel se place l'ensemble de ces travaux. La première est due à Bosq (1991) dans un article traitant de la modélisation des processus autorégressifs hilbertiens : *"These being nonparametric by themselves, it seems rather heavy to introduce a nonparametric model for observation lying in functional space ..."*. La seconde se trouve dans l'ouvrage de Ramsay et silverman (1997) "Functional Data analysis" dans le paragraphe intitulé "Challenges for the future" : *"theoretical aspects of Functional Data analysis have not been researched in sufficient depth, and it is hoped that appropriate theoretical developments will feed back into advances in practical methodology"*.
Ce document de synthèse s'inscrit pleinement dans le prolongement de ces réflexions tout en apportant de nouveaux (et parfois de premiers) éléments de réponse. Les travaux à venir découlant des nombreuses perspectives offertes par ce champ d'investigation ne manqueront pas de compléter et d'enrichir ceux déjà présents dans ce manuscrit.

Bibliographie

[1] C. Abraham, E. Mazner-Lϕber and M. Wegkamp, Universally consistent function classification, to appear in Scand. J. Statist.

[2] G. Aneiros Pérez, Esitmatión de modelos parcialmente lineales con errores dependientes, PhD Santiago de Compostela (2001).

[3] G. Aneiros Pérez, H. Cardot, G. Estévez Pérez and Ph. Vieu, Maximum ozone concentration forecasting by functional nonparametric approaches, Proceedings of TIES 2002 meeting, Genova Italia, 2002.

[4] A. Antoniadis, Autoregressive processes in Hilbert spaces, Contribution to the International Conference on Current Advances and Trends in Nonparametric Statistics, Crete July 15-19, 2002.

[5] A. Antoniadis and J. Beder, Joint estimation of the mean and the covariance of a banach valued gaussian vector, Ann. Stat., 20, 73-93 (1989).

[6] A. Berlinet and S. Levallois, Higher order analysis at Lebesgue points. In *Asymptotics in Statistics and Probability*, Ed. M. Puri, VSP.

[7] Ph. Besse, H. Cardot and F.Ferraty, Simultaneous nonparametric regressions of unbalanced longitudinal data, Computational Statistics & Data analysis, 24, 255-270 (1997).

[8] Ph. Besse, H. Cardot and D. Stephenson, Autoregressive forecasting of some functional climatic variations, Scandinavian Journal of Statistics, No. 27 673 687 (2000).

[9] D. Bosq, Nonparametric Statistics for Stochastic Processes : Estimation and Prediction (2nd edition), Lecture Notes in Statistics, 110, 1998.

[10] D. Bosq, Modelization, nonparametric estimation and prediction for continuous time processes, In Nonparametric functional estimation and related topics (Ed. G.G. Roussas), NATO ASI Series Kluwer Academic Publishers, 1991.

[11] D. Bosq, Limit theorems for Banch-valued autoregressive processes ; application to real continuous time processes,bull. Belg. Math. Soc., 9, 537-555 (1996).

[12] D. Bosq, Représentation autorégressive de l'opérateur de covariance empirique d'un ARH(1) ; applications, C.R. Acad. Sci. Paris, No 329, 531-534 (1999).

[13] D. Bosq, Linear Processes in Function Spaces : Theory and Applications, Lecture Notes in Statistics 149, Springer-Verlag, New-York, 2000.

[14] D. Bosq and M. Delecroix, Nonparametric prediction of a Hilbert space valued random variable, Stochatic Process Appl. No. 19 271 280 (1985).

[15] A. Boudou et Y. Romain, On spectral and random measures associated to continuous and discrete time processes, Statist. & Prob. Letters, 59, 145-157 (2002).

[16] J. Boularan, L. Ferré and P. Vieu, Location of particular points in nonparametric regression analysis, Australian J. of Statist., 37, 161-169 (1995).

[17] B. Brumback and J. Rice, Smoothing spline models for the analysis of nested and crossed samples of curves, J. Am. Statist. Assoc., 93, 961-976 (1998).

[18] C. Buell, Integral equation representation for factor analysis, Journal of Atmospheric Science, 28, 1502-1505 (1971).

[19] H., Cardot, F. Ferraty, A. Mas and P. Sarda, Testing Hypotheses in the Functional Linear Model, Scandinavian Journal of Statistics, 30, 241-255, 2003.

[20] H. Cardot, F. Ferraty and P. Sarda, Functional Linear Model, Statist. & Prob. Letters, 45, 11-22 (1999a).

[21] H. Cardot, F. Ferraty and P. Sarda, Spline estimators for the functional linear model : consistency, application and Splus implémentation, Rapport technique, BIA-INRIA No 1999/1 (1999b).

[22] H., Cardot, F. Ferraty and P. Sarda, Spline Estimators for the Functional Linear Model, to appear in Statistica Sinica, 2003.

[23] F. Chatelin, Spectral approximation of linear operators, Academic, New York, 1983.

[24] G. Collomb, Propriétés de convergence presque complète du prédicteur à noyau, Z. Wahrsch. verw. Geb. No. 66, 441 460 (1984).

[25] A. Cuevas, M. Febrero and R. Fraiman, Linear function regression : the case of fixed design and functional response, Canad. J. of Statistics, No. 30 285 300 (2002).

[26] J. Damon and S. Guillas, The inclusion of exogenous variables in functional autoregressive ozone forecasting, Environmetrics, No. 13 759 774 (2002).

[27] J. Dauxois, G.-M. Nkiet et Y. Romain, Projecteurs orthogonaux, opérateurs associés et Statistique multidimensionnelle (in french), Annales de l'ISUP, 45, fasc.1, 31-54 (2001).

[28] J. Dauxois et A. Pousse, Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique (in french), PhD Toulouse III (1976).

[29] J. Dauxois, A. Pousse and Y. Romain, Asymptotic theory for the principal component analysis of a random vector function : some applications to statistical inference, J. Multi. Anal., 12, 136-154 (1982).

[30] C. de Boor, A practical guide to splines, Springer, New York, 1978.

[31] M. Delecroix and M. Hristache, M-estimateur semi-paramétriques dans les modèles à direction révélatrice unique, Bull. Belg. Math. Soc., No 6, 161-185.

[32] J.C. Deville, Méthodes statistiques et numériques de l'analyse harmonique (in french), Ann. Insee, 15 (1974).

[33] L. Devroye and T.J. Wagner, Distribution-free consistency results in noparametric discrimination and regression function estimation, Ann. of Statist, 8 231 239 (1980).

[34] R.M. Dudley, Distances of probability measures and random variables, Ann. Statist., 39, 1563-1572, (1968).

[35] N. Dunford and J. Schwartz, Linear operators, Interscience Publishers, New York, 1963.

[36] J. Fan and I. Gijbels, Local polynomial modeling and its applications, Chapman and Hall, London, 1996.

[37] B. Fernández de Castro, S. Guillas and W. González Manteiga, Functional samples and bootstrap for the prediction of SO2 levels, Preprint (2003).

[38] F. Ferraty, A. Goia and Ph. Vieu, Functional model for time series : a fractal approach to dimension reduction. Test, No. 11 317 344 (2002a).

[39] F. Ferraty, A. Goia and Ph. Vieu, Statistica Funzionale : modelli di regressione non-parametrici (in italian), Collana Statistica, Franco-Angeli, Milano, 2002b.

[40] F. Ferraty, A. Goia and Ph. Vieu, Régression non-paramétrique pour des variables aléatoires fonctionnelles mélangeantes , C.R. Acad. Sci. Paris, No. 334 217-220 (2002c).

[41] F. Ferraty, A. Laksaci and Ph. Vieu, Estimating some characteristics of the conditional distribution in nonparametric functional models, Prepint (2003).

[42] F. Ferraty, V. Núñez Antón and Ph. Vieu, Regresión no paramétrica : desde la dimensión uno hasta la dimensión infinita (in spanish), Serv. Edit. Univers. País Vasco, Bilbao, 2001.

[43] F. Ferraty, A. Peuch and Ph. Vieu, Modèle à indice fonctionnel simple (in french), to appear in C.R. Acad. Sci. Paris (2003).

[44] F. Ferraty and Ph. Vieu, Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés (in french), C.R. Acad. Sci. Paris, No. 330 403 406 (2000).

[45] F. Ferraty and Ph. Vieu, Statistique Fonctionnelle : Modèles de régression pour variables aléatoires uni, multi et infiniment dimensionées (in french), Coursebook at Univ. P. Sabatier, Toulouse, France, Free access on line at http ://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html, 2001.

[46] F. Ferraty and Ph. Vieu, The functional nonparametric model and application to spectrometric data, Computational Statistics, No. 17 545 564 (2002).

[47] F. Ferraty and Ph. Vieu, Curves discrimination : a nonparametric functional approach, Computational Statistics & Data Analysis, In print, (2003a).

[48] F. Ferraty and Ph. Vieu, Nonparametric models for functional data, with applications in regression, time series prediction and curves discrimination. J. Nonparametric Statistics, In print (2003b).

[49] L.Ferré, J. Dauxois and A.-F. Yao, Un modèle semi-paramétrique pour variables aléatoires hilbertiennes (in french), C. R. Acad. Sci. Paris Sér. I Math. 333, no. 10, 947-952 (2001).

[50] I.E. Frank and J.H. Friedman, A Statistical View of Some Chemometrics Regression Tools, Technometrics, 35, 109-148 (1993).

[51] T. Gasser, P. Hall and B. Presnell, Nonparametric estimation of the mode of a distribution of random curves. J. R. Statist. Soc. B, No. 60 681 691 (1998).

[52] J. Geffroy, Estimation de la densité dans un espace métrique, C.R. Acad. Sci. Paris, No. 278 1149 1452 (1974).

[53] I.C. Gohberg and M.G. Krejn, Introduction à la théorie des opérateurs linéaires non auto-adjoints dans un espace hilbertien (in french), Dunod, Paris, 1971.

[54] A. Goia, Contribution à l'étude des modèles de régression pour variables aléatoires fonctionnelles (in french), PhD Toulouse III, 2003.

[55] G. Golubev and W. Härdle, On adaptive smoothing in partial linear models, Mathematical Methods of Statistics, No 11, 98-117 (2002).

[56] U. Grenander, Probabilities on algebraic structures, Almquist and Wilsell, 1963.

[57] P. Hall and N. Heckman, Estimating and depicting the structure of a distribution of random functions, Biometrika, No. 89 145 158 (2002).

[58] P. Hall, P. Poskitt and D. Presnell, A functional data-analytic approach to signal discrimination, Technometrics, No. 43 1 9 (2001).

[59] W. Härdle, P. Hall and H. Ichimura, Optimal smoothing in single-index models, Ann. of Statist., No. 21, 157-178 (1993).

[60] W. Härdle and J. Marron, Optimal bandwidth selection, Ann. of Statist., No. 13(4) 1465 1487 (1985).

[61] T. Hastie, A. Buja and R. Tibshirani, Penalized discriminant analysis. Ann. Statist., 23 73 102 (1995).

[62] T. Hastie and C. Mallows, Discuss. Technometrics No. 35 140 143 (1993).

[63] A.E. Hoerl and R.W. Kennard, Ridge regression : biased estimation for nonorthogonal problems, Technometrics, 8, 27-51 (1970).

[64] I. Holmström, On a method for parametric representation of the state of the atmosphere, Tellus, No 15, 127-149 (1963).

[65] D. Hoover, J. Rice, O. Wu and L. Yang, Nonparametric smoothing etimates of time-varying coefficient models with longitudinal data, Biometrika, 4, 809-822 (1998).

[66] P. Jacob and P. Oliveira, General approach to nonparametric histogram estimation, Statistics, No. 27 73 92 (1995).

[67] G.M. James, T.J. Hastie and C.A. Sugar, Principal component models for sparse functional data, Biometrika, No 87 587 602 (2000)

[68] T. Kato, Perturbation theory for linear operators, Springer, 1976.

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

[70] C. Kooperber, S. Bose and C.J. Stone, Polychotomous regression, J. Amer. Stat. Ass., No 92, 117-127 (1997).

[71] A. Kneip and T. Gasser, Statistical tools to analyse data representing a sample of curves, Ann. Statist., 20, 1266-1305 (1992).

[72] S.E Leurgans, R.A. Moyeed and B.W. Silverman, Canonical correlation analysis when the data are curves, J. R. Statist. Soc. B, 55, 725-740 (1993).

[73] S. Levallois, Convergence et normalité asymptotique d'estimateurs de la densité. Cas i.i.d. et ergodique, PhD Montpellier II, 1998.

[74] W.V. Li and Q.M. Shao, Gaussian processes : Inequalities, Small Ball Probabilites and Applications. In Stochastic processes : Theory and Methods, Ed. C.R. Rao and D. Shanbhag, 2001.

[75] R. Lipster and A. Shiryayev, On the absolute continuity of measures corresponding to processes of diffusion type relative to a Wiener measure. Izv Akad Nauk SSSR Ser Mat, No. 36 4 (1972).

[76] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T Zhang and K.L. Cohen, Robust component analysis for functional data (with discussion), No 8, 1 74 (1999).

[77] M. Loeve, Fonctions aléatoires du second ordre (in french), C.R. Acad. Sci. Paris, Série I, 220-469 (1945).

[78] Z. Luo, Backfitting in Smoothing Spline ANOVA, Ann. Stat., 26, 5, 1733-1759 (1998).

[79] Z. Luo, G. Wahba and D. Johnson, Spatial-Temporal Analysis of Temperature Using Smoothing Spline ANOVA, J. Climate, 11, 18-28 (1998).

[80] H. Martens and T. Naes, Multivariate Calibration, New York : John Wiley (1989).

[81] B. Marx and P. Eilers, Generalized linear regression on sampled signals and curves : a P-spline approach, Technometrics, 41, 1-13 (1999).

[82] A. Mas, Normalité asymptotique de l'estimateur empirique de l'opérateur d'autocorrélation d'un processus ARH(1) (in french), C.R. acad. Sci. Paris, No 329, 899-902 (1999).

[83] W.F. Massy, Principal components regression in exploratory statistical research, J. Amer. Stat. Soc., 60, 234-246 (1965).

[84] F. Merlevède, M. Peligrad and S. Utev, Sharp conditions for the CLT of linear processes in a Hilbert space, Journal of Theor. Probab., 10, 681-693 (1997).

[85] P. Molenaar and D. Boomsma, The genetic analysis of repeated measures : the karhunen-loeve expansion, Behavior Genetics, 17, 229-242 (1987).

[86] B. Nagy and F. Riesz, Leçons d'analyse fonctionnelle, Gauthier-Villars, Paris, 1965.

[87] S. Niang, Estimation de la densité en dimension infinie : Application aux processus de type diffusion (in french), C.R. Acad. Sci. Paris, No. 334 213 216 (2002a).

[88] S. Niang, Sur l'estimation de la densité en dimension infinie : application aux diffusions (in french), PhD Paris VI, 2002b.

[89] V. Nunez-Antón, J. Rodriguez-Póo et P. Vieu (1999). Longitudinal data with nonstationnary errors : a nonparametric three stage approach, TEST, 8, 201-231 (1999).

[90] A. Obhukov, the statistically orthogonal expansion of empirical functions, American Geophysical Union, 288-291 (1960).

[91] Parthasarathy, Probability Measures on metric spaces, Academic Press, 1967.

[92] Ya.B. Pesin, On rigourous mathematical definitions of correlation dimension and generalized spectrum for dimensions, J. Statis. Phys., 71 529 547 (1993)

[93] S. Pezzulli and B. Silverman, On smoothed principal components analysis, Computional Statistics, 8, 1-16 (1993).

[94] C. Preda, Approximation par moyennage de l'analyse en composantes principales d'un processus stochastique (in french), C.R. Acad. Sci. Paris, No. 330 605-610 (2000).

[95] C. Preda, Analyse factorielle d'un processus : problèmes d'approximation et de régression (in french), PhD Lille I, 1999.

[96] B. Pumo, Prediction of continuous time processes by $C[0,1]$-valued autoregressive process, Statist. Infer. for Stoch. Processes, 3, 1-13 (1999).

[97] J. Ramsay, Principal differential analysis : data reduction by differential operators, J. R. Stat. Soc. B, 58, 495-508 (1996).

[98] J. Ramsay, N. Altman and R.D. Bock, Varaition in height acceleration in the fels growth data, Canad. J. Statist., 22, 89-102 (1994).

[99] J. Ramsay, X. wang and R. Flanagan, A functional data analysis of the pinch force of human fingers, Applied statist., 44, 17-30 (1995)

[100] J. Ramsay and C. Dalzell, Some tools for functional data analysis, J. R. Statist. Soc. B, 53, 539-572 (1991).

[101] J. Ramsay and X. Li, Curve registration, J. R. Statist. Soc. B, 60, 351-363 (1996).

[102] J. Ramsay and B. Silverman, Functional Data Analysis, Springer-Verlag, New-York, 1997.

[103] J. Ramsay and B. Silverman, Applied functional data analysis : Methods and case studies, Springer-Verlag, New York, 2002.

[104] N. Rhomari, Kernel regression in Banach space, Contribution to the International Conference on Current Advances and Trends in Nonparametric Statistics, Crete July 15-19, 2002.

[105] J. Rice and B. Silverman, Estimating the mean and the covariance structure nonparametrically when the data are curves, J. R. Statist. Soc. B, 53, 233-243 (1991).

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

[107] E. Rio, Théorie asymptotique des processus faiblement dépendants (in french), SMAI Math & Applications 31 Springer, 1999.

[108] Y. Romain, Étude asymptotique des approximations par échantillonnage de l'analyse en composantes principales d'une fonction aléatoire. Quelques applications, PhD Toulouse III, 1979.

[109] Y. Romain, Contributions la Statistique operatorielle. Habilitation diriger des recherches, Toulouse III, 2000.

[110] Y. Romain, Perturbation of functional tensors with applications to covariance operators, Statist. & Proba. Lett. No 58, 253-264, (2002).

[111] G. Roussas, Nonparametric estimation of the transition distribution function of a Markov process, Ann. Math. Statist. No. 40 1386 1400 (1969).

[112] M. Samanta, Non-parametric estimation of conditional quantiles, Statist. & Proba. Lett. No. 7, 407-412 (1989).

[113] G. Saporta, Méthodes exploratoires d'analyse des données temporelles (in french), Cahiers du B.U.R.O., No 37-38, Université Pierre et Marie Curie.

[114] M. Schimek (ed.), Smoothing and regression : Approaches, Computation, and Application, Wiley Series in Probability and Statistics, Wiley, New-York, 2000.

[115] B.W. Silverman, Incorporating effects into functional principal components analysis, J. R. Stat. Soc. B, 57, 673-689 (1995).

[116] B.W. Silverman, Smoothed functional principal components analysis by choice of norm, Ann. Stat., 24, 1-24 (1996).

[117] Staph, Birth and first activities of the working group in Functional Statistics, Tech. Report (see http ://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html), Univ. P. Sabatier, Toulouse, France, No. 05 (2001).

[118] Staph, Proceedings of the working group in Functional and Operatorial Statistics, Tech. Report (see http ://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html), Univ. P. Sabatier, Toulouse, France, No. 12 (2002).

[119] J.V. Uspensky, Introduction to mathematical probability, Mc Graw-Hill Book Company, Inc., New York and London, 1937.

[120] Y. Wang, Mixed effects smoothing spline analysis of variance, J. R. Statist. Soc. B No. 60 159 174 (1998).

[121] W. Whitt, Preservation of rates of convergence under mappings, Z. Wahr. verw. Gebiete, 29, 39-44 (1974).

[122] H. Wold, Soft Modelling by latent variables ; the nonlinear iterative partial least squares approach, In Perspectives in Probability and Statistics, Papers in Honour of M.S. Barlett, ed. J. Giani, London : Academic Press (1975).

[123] V.V. Yurinskii, Exponential Inequalities for sums of random vectors, Journal of Multivarite Analysis, 6, 473-499 (1976)