Contributions to Statistics

Frédéric Ferraty *Editor*

Recent Advances in Functional Data Analysis and Related Topics

Physica-Verlag

Contributions to Statistics

For other titles published in this series, go to
www.springer.com/series/2912

Frédéric Ferraty
Editor

Recent Advances in Functional Data Analysis and Related Topics

Physica-Verlag
A Springer Company

Editor
Frédéric Ferraty
Mathematics Toulouse Institute
Toulouse University
Narbonne road 118
31062 Toulouse
France
frederic.ferraty@math.univ-toulouse.fr

ISSN 1431-1968
ISBN 978-3-7908-2735-4 e-ISBN 978-3-7908-2736-1
DOI 10.1007/978-3-7908-2736-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011929779

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

Physica-Verlag is a brand of Springer-Verlag Berlin Heidelberg
Springer -Verlag is part of Springer Science+Business Media (www.springer.com)

Preface

Nowaday, the progress of high-technologies allow us to handle increasingly large datasets. These massive datasets are usually called ”high-dimensional data” . At the same time, different ways of introducing some continuum in the data appeared (use of sophisticated monitoring devices, function-based descriptors as the density function for instance, etc). Hence, the data can be considered as observations varying over a continuum defining a subcategory of high-dimensional data called functional data. Statistical methodologies dealing with functional data are called Functional Data Analysis (FDA), the ”functional” word emphasizing the fact that the statistical method takes into accound the functional feature of the data. The failure of standard multivariate statistical analyses, the numerous fields of applications as well as the new theoretical challenges motivate an increasingly statistician community to develop new methodologies. The huge research activity around FDA and its related fields produces very fast progress. Then, it is necessary to propose regular snapshots about the most recent advances in this topic.

This is the main goal of the International Workshop on Functional and Operatorial Statistics (IWFOS’2011, Santander, Spain) which is the second edition of the first successful one (IWFOS’2008, Toulouse, France) initiated by the working group STAPH (Toulouse Mathematics Institute, France). This volume gathers peer-reviewed contributions authored by outstanding confirmed experts as well as young brillant researchers. The presentation of these contributions in a short (around 6 pages a contribution) and concise way makes the reading and use of this book very easy. As a by-product, the reader should find most of representative and significant recent advances in this field, mixing works oriented towards applications (with original datasets, computational issues, applications in numerous fields of Sciences - biometrics, chemometrics, economics, medicine, etc) with fundamental theoretical ones. This volume contents a wide scope of statistical topics: change point detection, clustering, conditional density/expectation/mode/quantiles/extreme quantiles, covariance operators, depth, forecasting, functional additive regression, functional extremality, functional linear regression, functional principal components analyses, functional single index model, functional varying coefficient models, generalized additive models, hilbertian processes, nonparametric models, noisy obser-

vations, quantiles in functions spaces, random fields, semi-functional models, statistical inference, structural tests, threshold-based procedures, time series, variable selection, wavelet-based smoothing, etc. These statistical advances deal with numerous kind of interesting datasets (functional data, high-dimensional data, longitudinal functional data, multidimensional curves, spatial functional data, sparse functional data, spatial-temporal data) and propose very attractive applications in various fields of Sciences: DNA minicircles, electoral behavior, electricity spot markets, electro-cardiogram records, gene expression, irradiance data (exploitation of solar energy), magnetic resonance spectroscopy data (neurocognitive impairment), material sciences, signature recognition, spectrometric curves (chemometrics), tractography data (multiple sclerosis), etc.

Clearly, this volume should be very attractive for a large audience, like academic researchers, graduate/PhD students as well as engineers using regularly recent statistical developments in his work.

At last, this volume is a by-product of the organization of IWFOS'2011 which is chaired by two other colleagues: Juan A. Cuesta-Albertos (Santander, Spain) and Wenceslao Gonzalez-Manteiga (Santiago de Compostela, Spain). Their trojan work as well as their permanent support and enthusiasm are warmly and gratefully thanked.

Toulouse, France
March 2011

Frédéric Ferraty
The Editor and co-Chair of IWFOS'2011

Acknowledgements

First of all, the vital material of this volume was provided by the contributors. Their outstanding expertise in this statistical area as well as their valuable contributions guarantee the high scientific level of this book and hence the scientific success of IWFOS'2011. All the contributors are warmly thanked.

This volume could not have existed without the precious and efficience help of the members of the IWFOS'2011 Scientific Committee named J. Antoch (Prague, Czech Rep.), E. del Barrio (Valladolid, Spain), G. Boente (Buenos Aires, Argentina), C. Crambes (Montpellier, France), A. Cuevas (Madrid, Spain), L. Delsol (Orléans, France), D. Politis (California, USA), M. Febrero-Bande (Santiago de Compostela, Spain), K. Gustafson (Colorado, USA), P. Hall (Melbourne, Ausralia), S. Marron (North Carolina, USA), P. Sarda (Toulouse, France), M. Valderrama (Granada, Spain), S. Viguier-Pla (Toulouse, France), Q. Yao (London, UK). Their helpful and careful involvement in the peer-reviewing process has contributed significantly to the high scientific level of this book; all of them are gratefully acknowledged.

Of course, this book is a by-product of IWFOS'2011 and its success is due to the fruitful collaboration of people from the University of Cantabria (Santander, Spain), the University of Santiago de Compostela (Spain) and the University of Toulouse (France). In particular, A. Boudou (Toulouse, France), A. Martínez Calvo (Santiago de Compostela, Spain), A. Nieto-Reyes (Santander, Spain), B. Pateiro-López (Santiago de Compostela), Gema R. Quintana-Portilla (Santander, Spain), Y. Romain (Toulouse, France) and P. Vieu (Toulouse, France), members of the Organizing Committee, have greatly contributed to the high quality of IWFOS'2011 and are gratefully thanked. It is worth noting that this scientific event is the opportunity to emphasize the links existing between these three universities and this is why these International Workshop is chaired by three people, one of each above-mentioned university. Clearly, this Workshop should strengthen the scientific collaborations between these three universities.

A special thank is addressed to the working group STAPH (http://www.math.univ-toulouse.fr/staph). Its intensive and dynamic research activities oriented towards functional and operatorial statistics with a special attention to Functional Data Anal-

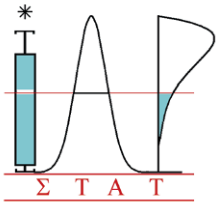
ysis and High-Dimensional Data contributed to the development of numerous scientific collaborations with statisticians in the whole world. A first consequence was the creation, the organization and the management of the first edition of IWFOS (Toulouse, France, 2008). The success of IWFOS'2008 was certainly the necessary starting point allowing the emergence of IWFOS'2011. All its members and collaborators are warmly acknowledged.

The final thanks go to institutions/organizations which supported this Workshop via grants or administrative supports. In particular, the Chairs of IWFOS'2011 would like to express their grateful thanks to:

- the Departamento de Matemáticas, Estadística y Computación, the Facultad de Ciencias and the Vicerrectorado de Investigación y Transferencia del Conocimiento de la Universidad de Cantabria,
- the Programa Ingenio Mathematica, iMATH,
- the Acciones Complementarias del Ministerio Español de Ciencia e Innovación,
- the Toulouse Mathematics Institute,
- the IAP research network in statistics

March 2011

Juan A. Cuesta-Albertos
Frédéric Ferraty
Wenceslao Gonzalez-Manteiga
The co-Chairs of IWFOS'2011



Contents

Chapter 1
Penalized Spline Approaches for Functional Principal Component Logit Regression

A. Aguilera, M. C. Aguilera-Morillo, M. Escabias, M. Valderrama

Abstract The problem of multicollinearity associated with the estimation of a functional logit model can be solved by using as predictor variables a set of functional principal components. The functional parameter estimated by functional principal component logit regression is often unsmooth. To solve this problem we propose two penalized estimations of the functional logit model based on smoothing functional PCA using P-splines.

1.1 Introduction

The aim of the functional logit model is to predict a binary response variable from a functional predictor and also to interpret the relationship between the response and the predictor variables. To reduce the infinite dimension of the functional predictor and solve the multicollinearity problem associated to the estimation of the functional logit model, Escabias et al. (2004) proposed to use a reduced number of functional principal components (pc's) as predictor variables. A functional PLS based solution was also proposed by Escabias et al. (2006). The problem associated with these approaches is that in many cases the estimated functional parameter is not smooth and therefore difficult to interpret. Different penalized likelihood estimations with B-spline basis were proposed in the general context of functional generalized linear

Ana Aguilera
Department of Statistics and O. R. University of Granada, Spain, e-mail: aaguiler@ugr.es

Maria del Carmen Aguilera-Morillo
Department of Statistics and O. R. University of Granada, Spain, e-mail: caguilera@ugr.es

Manuel Escabias
Department of Statistics and O. R. University of Granada, Spain, e-mail: escabias@ugr.es

Mariano Valderrama
Department of Statistics and O. R. University of Granada, Spain, e-mail: valderra@ugr.es

F. Ferraty (ed.), *Recent Advances in Functional Data Analysis and Related Topics*,
DOI 10.1007/978-3-7908-2736-1_1,

models to solve this problem (Marx and Eilers, 1999; Cardot and Sarda, 2005). In this paper we introduce two different penalized estimation approaches based on smoothed functional principal component analysis (FPCA). On one hand, FPCA of P-spline approximation of sample curves is performed. On the other hand, a discrete P-spline penalty is included in the own formulation of FPCA.

1.2 Background

Let us consider a sample of functional observations $x_1(t), x_2(t), \ldots, x_n(t)$ of a fixed design functional variable and let $y_1, y_2, \ldots, y_n$ be a random sample of a binary response variable Y associated to them. That is, $y_i \in \{0,1\}$, $i = 1, \ldots, n$. The functional logistic regression model is given by

$$y_i = \pi_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where π_i is the expectation of Y given $x_i(t)$ modeled as

$$\pi_i = P[Y = 1 | \{x_i(t) : t \in T\}] = \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)\,dt\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)\,dt\}}, \quad i = 1, \ldots, n,$$

α being a real parameter, $\beta(t)$ a parameter function, $\{\varepsilon_i : i = 1, \ldots, n\}$ independent errors with zero mean and T the support of the sample paths $x_i(t)$.

The logit transformations can be expressed as

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \int_T x_i(t)\beta(t)\,dt, \quad i = 1, \ldots, n. \tag{1.1}$$

A way to estimate the functional logit model is to consider that both, the sample curves and the parameter function, admit an expansion in terms of basis functions. Then, the functional logit model turns into a multiple logit model whose design matrix is the product between the matrix of basis coefficients of sample paths and the matrix of inner products between basis functions (Escabias et al., 2004). The estimation of this model is affected by multicollinearity due to the high correlation between the columns of the design matrix. In order to obtain a more accurate and smoother estimation of the functional parameter than the one provided by standard functional principal component logit regression (FPCLR), we present in this paper two penalized estimation approaches based on P-spline smoothing of functional PCA.

1.3 Penalized estimation of FPCLR

In general, the functional logit model can be rewritten in terms of functional principal components as

$$L = \alpha \mathbf{1} + \Gamma \gamma, \tag{1.2}$$

where $\Gamma = (\xi_{ij})_{n\times p}$ is a matrix of functional pc's of $x_1(t),\ldots,x_n(t)$ and γ is the vector of coefficients of the model.

By considering that the predictor sample curves admit the basis expansions $x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t)$, the functional parameter can be also expressed also in terms of the same basis, $\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k(t)$, and the vector β of basis coefficients is given by $\hat{\beta} = F\hat{\gamma}$, where the way of computing F depends on the kind of FPCA used to obtain the pc's.

An accurate estimation of the parameter function can be obtained by considering only a set of optimal principal components as predictor variables. In this paper we select the optimal number of predictor pc's by using a leave-one-out cross validation method that maximizes the area under ROC curve computed by following the process outlined in Mason and Graham (2002). To obtain this area, observed and predicted values are required. In this case, we have considered y_i the i^{th} observed value of the binary response and $\hat{y}_i^{(-i)}$ the i^{th} predicted value obtained by deleting the i^{th} observation of the design matrix in the iterative estimation process.

Let us consider that the sample curves are centered and belong to the space $L^2[T]$ with the usual inner product defined by $< f,g >= \int_T f(t)g(f)dt$. In the standard formulation of functional PCA, the j^{th} principal component scores are given by

$$\xi_{ij} = \int_T x_i(t) f_j(t)\, dt,\ i = 1,\ldots,n, \tag{1.3}$$

where the weight function or factor loading f_j is obtained by solving

$$\begin{cases} Max_f\ \ Var\left[\int_T x_i(t) f(t)\, dt\right] \\ s.t.\ \|f\|^2 = 1\ and\ \int f_\ell(t) f(t)\, dt = 0,\ \ \ell = 1,\ldots,j-1. \end{cases}$$

The weight functions f_j are the solutions to the eigenequation $Cf_j = \lambda_j f_j$, with $\lambda_j = Var[\xi_j]$ and C being the sample covariance operator defined by $Cf = \int c(.,t) f(t)\, dt$, in terms of the sample covariance function $c(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s) x_i(t)$.

In practice, functional PCA has to be estimated from discrete time observations of each sample curve $x_i(t)$ at a set of times $\{t_{i0}, t_{i1},\ldots,t_{im_i} \in T,\ i = 1,\ldots,n\}$. The sample information is given by the vectors $x_i = (x_{i0},\ldots,x_{im_i})'$, with x_{ik} the observed value for the i^{th} sample path $x_i(t)$ at time t_{ik} $(k = 0,\ldots,m_i)$.

When the sample curves are smooth and observed with error, least squares approximation in terms of B-spline basis is an appropriate solution for the problem of reconstructing their true functional form. This way, the vector of basis coefficients of each sample curve that minimizes the least squares error is given by $\hat{a}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i' x_i$, with $\Phi_i = (\phi_j(t_{ik}))_{m_i\times p}$ and $a_i = (a_{i1},\ldots,a_{ip})'$.

Functional PCA is then equivalent to the multivariate PCA of $A\Psi^{\frac{1}{2}}$ matrix, $\Psi^{\frac{1}{2}}$ being the squared root of the matrix of the inner products between B-spline basis functions (Ocaña et al. 2007). Then, matrix F that provides the relation between the basis coefficients of the functional parameter and the parameters in terms of principal components is given by $F = \Psi_{p\times p}^{-\frac{1}{2}} G_{p\times n}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$. This non smoothed FPCA estimation of functional logit models with B-spline basis was performed by Escabias et al. (2004).

1.3.1 Functional PCA via P-splines

Now, we propose a penalized estimation based on functional PCA of the P-spline approximation of the sample curves. The basis coefficients in terms of B-splines are computed by introducing a discrete penalty in the least squares criterion (Eilers and Marx, 1996), so that we have to minimize $(x_i - \Phi_i a_i)'(x_i - \Phi_i a_i) + \lambda a_i' P_d a_i$, where $P_d = \left(\triangle^d\right)' \triangle^d$ and $\triangle^d$ is the differencing matrix that gives the dth-order differences of a_i. The solution is then given by $\hat{a}_i = \left(\Phi_i'\Phi_i + \lambda P_d\right)^{-1} \Phi_i' x_i$, and the smoothed parameter is chosen by leave-one-out cross validation.

Then, we carry out the multivariate PCA of $A\Psi^{\frac{1}{2}}$ matrix as explained above. The difference between smoothed FPCA via P-splines and non smoothed FPCA is only the way of computing the basis coefficients (rows of matrix A), with or without penalization, respectively.

1.3.2 P-spline smoothing of functional PCA

Now we propose to obtain the principal components by maximizing a penalized sample variance that introduces a discrete penalty in the basis coefficients of principal component weights.

The j^{th} principal component scores are defined as in equation (1.3) but now the weight functions f_j are obtained by solving

$$\begin{cases} Max_f \ \dfrac{var[\int x_i(t) f(t)\, dt]}{\|f\|^2 + \lambda PEN_d(f)} \\ s.t.\ \|f\|^2 = b'\Psi b = 1 \ and \ b'\Psi b_l + b' P_d b_l = 0,\ \ell = 1,\ldots,j-1, \end{cases}$$

where $PEN_d(f) = b' P_d b$ is the discrete roughness penalty function, b being the vector of basis coefficients of the weight functions, $f(t) = \sum_{k=1}^{p} b_k \phi_k$, and λ the smoothing parameter estimated by leave-one-out cross validation.

Finally, this variance maximization problem is converted into an eigenvalue problem, so that, applying the Choleski factorization $LL' = \Psi + \lambda P_d$, P-spline smooth-

ing of functional PCA is reduced to classic PCA of the matrix $A\Psi(L^{-1})'$. Then, the estimated vector β of basis coefficients of the functional parameter is given by $\hat{\beta} = F\hat{\gamma} = \left(L^{-1}\right)' G\hat{\gamma}$, where G is the matrix of eigenvectors of the sample covariance matrix of $A\Psi(L^{-1})'$.

1.4 Simulation study

We are going to illustrate the good performance of the proposed penalty approaches following the simulation scheme developed in Ferraty and Vieu (2003) and Escabias et al. (2006). We simulated 1000 curves of two different classes of sample curves. For the first class we simulated 500 curves according to the random function $x(t) = uh_1(t) + (1-u)h_2(t) + \varepsilon(t)$, and another 500 curves were simulated for the second class according to the random function $x(t) = uh_1(t) + (1-u)h_3(t) + \varepsilon(t)$, u and $\varepsilon(t)$ being uniform and standard normal simulated random values, respectively, and $h_1(t) = \max\{6 - |t-11|, 0\}, h_2(t) = h_1(t-4), h_3(t) = h_1(t+4)$. The sample curves were simulated at 101 equally spaced points in the interval $[1,21]$.

As binary response variable, we considered $Y = 0$ for the curves of the first class and $Y = 1$ for the ones of the second class. After simulating the data, we performed least squares approximation of the curves, with and without penalization, in terms of the cubic B-spline functions defined on 30 equally spaced knots of the interval $[1,21]$.

	non smoothed FPCA	FPCA via P-splines	P-spline smoothed FPCA
Number pc's	3	2	3
ROC area	0.9986	0.9985	0.9988

Table 1.1: Area under the ROC curve for the test sample with the optimum models selected by cross validation with the three different FPCA approaches (non smoothed FPCA, FPCA via P-splines ($\lambda = 24.2$) and P-spline smoothed FPCA ($\lambda = 5$)).

In order to estimate the binary response Y from the functional predictor X we have estimated three different FPCLR models by using non smoothed FPCA and the two P-spline estimation approaches of FPCA proposed in this work. A training sample of 500 curves (250 of each class) was considered to fit the model and a test sample with the remaining 500 curves to evaluate the forecasting performance of the model. The pc's were included in the model by variability order and the optimum number of pc's selected by maximizing the cross validation estimation of the area under the ROC curve. In Table 1.1 we can see that P-spline smoothed FPCA estimation provides a slightly higher area and FPCA via P-splines requires fewer components.

Escabias et al. (2006) estimated the parameter function using different methods as functional PLS logistic regression and functional principal component logit regression, obtaining in both cases a non smooth estimation. In Figure 1.1 we can see that both penalized estimations of FPCA based on P-splines provide a smooth estimation of the functional parameter. This shows that using a smoothing estimation of FPCA is required in order to obtain a smooth estimation of the functional parameter that makes the interpretation easier. Although there are not significant differences between the estimation of the parameter function provided by FPCA via P-splines and P-spline smoothed FPCA, the second approach spends much more time in cross validation procedure so that, in practice, the estimation of FPCLR based on FPCA via P-splines is more efficient.

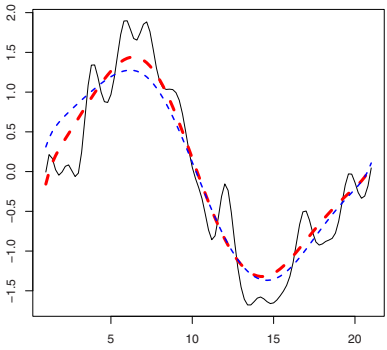


Fig. 1.1: Estimated parameter function with the three different considered FPCA estimations: non smoothed FPCA (black and continue line), FPCA via P-splines (red and long dashed line, $\lambda = 24.2$) and P-spline smoothed FPCA (blue and short dashed line, $\lambda = 5$)

Acknowledgements This research has been funded by project MTM2010-20502 from *Ministerio de Ciencia e Innovación, Spain.*

References

1. Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**(2), 89–121 (1996)
2. Cardot, H., Sarda, P.: Estimation in generalized linear models for functional data via penalized likelihood. J. Multivariate Anal. **92**(1), 24–41 (2005)

3. Escabias, M., Aguilera, A. M., Valderrama. M. J.: Principal component estimation of functional logistic regression: discussion of two different approaches. J. Nonparametr. Stat. **16**(3-4), 365–384 (2004)
4. Escabias, M., Aguilera, A. M., Valderrama. M. J.: Functional PLS logit regression model. Comput. Stat. Data An. **51**, 4891–4902 (2006)
5. Ferraty, F., Vieu, P.: Curves discrimination: a nonparametric functional approach. Comput. Stat. Data An. **44**, 161–173 (2003)
6. Marx, B.D., Eilers, P.H.C.: Generalized linear regression on sampled signals and curves. A P-spline approach. Technometrics **41**, 1–13 (1999)
7. Mason, S.J., Graham, N.E.: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q. J. Roy. Meteor. Soc. **128**, 291–303 (2002)
8. Ocaña, F.A., Aguilera, A.M. and Escabias, M.: Computational considerations in functional principal component analysis. Computation. Stat. **22**(3), 449–465 (2007)

Chapter 2
Functional Prediction for the Residual Demand in Electricity Spot Markets

Germán Aneiros, Ricardo Cao, Juan M. Vilar-Fernández, Antonio Muñoz-San-Roque

Abstract The problem of residual demand prediction in electricity spot markets is considered in this paper. Hourly residual demand curves are predicted using non-parametric regression with functional explanatory and functional response variables. Semi-functional partial linear models are also used in this context. Forecasted values of wind energy as well as hourly price and demand are considered as linear predictors. Results from the electricity market of mainland Spain are reported. The new forecasting functional methods are compared with a naive approach.

2.1 Introduction

Nowadays, in many countries all over the world, the production and sale of electricity is traded under competitive rules in free markets. The agents involved in this market: system operators, market operators, regulatory agencies, producers, consumers and retailers have a great interest in the study of electricity load and price. Since electricity cannot be stored, the demand must be satisfied instantaneously and producers need to anticipate to future demands to avoid overproduction. Good forecasting of electricity demand is then very important from the system operator viewpoint. In the past, demand was predicted in centralized markets (see Gross and Galiana (1987)) but competition has opened a new field of study. On the other hand

Germán Aneiros
Universidade da Coruña, Spain, e-mail: ganeiros@udc.es

Ricardo Cao
Universidade da Coruña, Spain, e-mail: rcao@udc.es

Juan M. Vilar-Fernández
Universidade da Coruña, Spain, e-mail: eijvilar@udc.es

Antonio Muñoz-San-Roque
Universidad Pontificia de Comillas, Madrid, Spain, e-mail: antonio.munoz@iit.icai.upcomillas.es

F. Ferraty (ed.), *Recent Advances in Functional Data Analysis and Related Topics*,
DOI 10.1007/978-3-7908-2736-1_2,

prediction of residual demand of an agent is a valuable tool to establish good bidding strategies for the agent itself. Consequently, prediction of electricity residual demand is a significant problem in this sector.

Residual demand curves have been considered previously in the literature. In each hourly auction, the residual demand curve is defined as the difference of the combined effect of the demand at any possible price and the supply of the generation companies as a function of price. Consequently 24 hourly residual demand curves are obtained every day. These curves are useful tools to design optimal offers for companies operating in a day-ahead market (see Baillo et al. (2004) and Xu and Baldick (2007)). We focus on one day ahead forecasting of electricity residual demand curves. Therefore, for each day of the week, 24 curve forecasts need to be computed.

This paper proposes functional and semi-functional nonparametric and partial linear models to forecast electricity residual demand curves. Forecasted wind energy as well as forecasted hourly price and demand are incorporated as explanatory variables in the model. Nonparametric regression estimation under dependence is a useful tool for time series forecasting. Some relevant work in this field include Härdle and Vieu (1992), Hart (1996) and Härdle, Lütkepohl and Chen (1997). Other papers more specifically focused on prediction using nonparametric techniques are Carbon and Delecroix (1993), Matzner-Lober, Gannoun and De Gooijeret (1998) and Vilar-Fernández and Cao (2007). The literature on methods for time series prediction in the context of functional data is much more limited. The books by Bosq (2000) and Ferraty and Vieu (2006) are comprehensive references for linear and nonparametric functional data analysis, respectively. Faraway (1997) considered a linear model with functional response in a regression setup. Antoch et al. (2008) also used functional linear regression models to predict electricity consumption. Antoniadis, Paparoditis and Sapatinas (2006) proposed a functional wavelet-kernel approach for time series prediction and Antoniadis, Paparoditis and Sapatinas (2009) studied a method for smoothing parameter selection in this context. Aneiros-Pérez and Vieu (2008) have dealt with the problem of nonparametric time series prediction using a semi-functional partial linear model and Aneiros-Pérez, Cao and Vilar-Fernández (2010) used Nadaraya-Watson and local linear methods for functional explanatory variables and scalar response in time series prediction. Finally, Cardot, Dessertaine and Josserand (2010) use semi-parametric models for predicting electricity consumption and Vilar-Fernández, Cao and Aneiros (2010) use also semi-functional models with scalar response to predict next-day electricity demand and price.

The remaining of this paper is organized as follows. In Section 2, a mathematical description of the functional nonparametric model is given. The semi-functional partial linear model is presented in Section 3. Section 4 contains some information about the data and the empirical study concerning one-day ahead forecasting of electricity residual demand curves in Spain. The references are included at the final section of the paper.

2.2 Functional nonparametric model

The time series under study (residual demand curve) will be considered as a realization of a discrete time functional valued stochastic process, $\{\chi_t(p)\}_{t\in\mathbb{Z}}$, observed for $p \in [a,b)$. For a given hour, r, ($r \in \{1,\ldots,24\}$) of day t, the values of $\chi_t^{(r)}(p)$ indicate the energy that can be sold (positive values) or bought (negative values) at price p and the interval $[a,b)$ is the range for prices. We first concentrate on predicting the curve $\chi_{n+1}^{(r)}(p)$, after having observed a sample of values $\left\{\chi_i^{(r)}(p)\right\}_{i=1,2,\ldots,n}$. For simplicity the superindex r will be dropped off.

In the following we will assume that the sequence of functional valued random variables $\{\chi_t(p)\}_{t\in\mathbb{Z}}$ is Markovian. We may look at the problem of predicting the future curve $\chi_{n+1}(p)$ by computing nonparametric estimations, $\widehat{m}(\chi)$, of the autoregression function in the functional nonparametric (FNP) model

$$\chi_{i+1}(\bullet) = m(\chi_i) + \varepsilon_{i+1}(\bullet),\ i = 1,\ldots,n, \tag{2.1}$$

which states that the values of the residual demand at day $i+1$ is an unknown nonparametric function of the residual demand at the previous day plus some error term. These errors $\varepsilon_i(\bullet)$ are iid zero mean functional valued random variables. Thus, $\widehat{m}(\chi_n)$ gives a functional forecast for $\chi_{n+1}(\bullet)$.

In our context this approach consists on estimating the autoregression functional, m, using hourly residual demand curves and apply this estimated functional to the last observed day.

Whereas the Euclidean norm is a standard distance measure in finite dimensional spaces, the notion of semi-norm or semi-metric arises in this infinite-dimensional functional setup. Let us denote by $\mathscr{H} = \{f : C \to \mathbb{R}\}$ the space where the functional data live and by $d(\bullet,\bullet)$ a semi-metric associated with $\mathscr{H}$. Thus $(\mathscr{H},d)$ is a semi-metric space (see Ferraty and Vieu (2006) for details).

A Nadaraya-Watson type estimator (see Nadaraya (1964) and Watson (1964)) for m in (2.1) is defined as follows

$$\widehat{m}_h^{FNP}(\chi) = \sum_{i=1}^{n-1} w_h(\chi,\chi_i)\chi_{i+1}(\bullet), \tag{2.2}$$

where the bandwidth $h > 0$ is a smoothing parameter,

$$w_h(\chi,\chi_i) = \frac{K(d(\chi,\chi_i)/h)}{\sum_{j=1}^{n} K(d(\chi,\chi_j)/h)}, \tag{2.3}$$

and the kernel function $K : [0,\infty) \to [0,\infty)$ is typically a probability density function chosen by the user.

The choice of the kernel function is of secondary importance. However, both the bandwidth and the semi-metric are relevant aspects for the good asymptotic and practical behavior of (2.2).

A key role of the semi-metric is that related to the so called "curse of dimensionality". From a practical point of view the "curse of dimensionality" can be explained as the sparsness of data in the observation region as the dimension of the data space grows. This problem is specially dramatic in the infinite-dimensional context of functional data. More specifically, Ferraty and Vieu (2006) have proven that it is possible to construct a semi-metric in such a way that the rate of convergence of the nonparametric estimator in the functional setting is similar to that of the finite-dimensional one. It is important to remark that we use a semi-metric rather than a metric. Indeed, the "curse of dimensionality" would appear if a metric were used instead of a semi-metric.

In functional data it is usual to consider semi-metrics based on semi-norms. Thus, Ferraty and Vieu (2006) recommend, for smooth functional data, to take as semi-norm the L_2 norm of some q-th derivative of the function. For the case of rough data curves, these authors suggest to construct a semi-norm based on the first q functional principal components of the data curves.

2.3 Semi-functional partial linear model

Very often there exist exogenous scalar variables that may be useful to improve the forecast. For the residual demand prediction this may be the case of the hourly wind energy in the market and the hourly price and demand. Although these values cannot be observed in advance, one-day ahead forecasts can be used to anticipate the values of these three explanatory variables. Previous experience also suggests that an additive linear effect of these variables on the values to forecast might occur. In such setups, it seems natural to generalize model (2.1) by incorporating a linear component. This gives the semi-functional partial linear (SFPL) model:

$$\chi_{i+1}(\bullet) = \mathbf{x}_{i+1}^T \beta(\bullet) + m(\chi_i) + \varepsilon_{i+1}(\bullet),\ i = 1,\ldots,n, \tag{2.4}$$

where $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^T \in \mathbb{R}^p$ is a vector of exogenous scalar covariates and $\beta(\bullet) = (\beta_1(\bullet),\ldots,\beta_p(\bullet))^T$ is a vector of unknown functions to be estimated.

Now, based on the SFLP model, we may look at the problem of predicting $\chi_{n+1}(\bullet)$ by computing estimations $\widehat{\beta}$ and $\widehat{m}(\chi)$ of β and $m(\chi)$ in (2.4), respectively. Thus, $\mathbf{x}_{n+1}^T \widehat{\beta}(\bullet) + \widehat{m}(\chi_n)$ gives the forecast for $\chi_{n+1}(\bullet)$.

An estimator for $\beta(\bullet)$ based on kernel and ordinary least squares ideas was proposed in Aneiros-Pérez and Vieu (2006) in the setting of independent data. More specifically, recall the weights $w_h(\chi,\chi_i)$ defined in the previous subsection and denote $\widetilde{\mathbf{X}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{X}$ and $\widetilde{\chi}_h = (\mathbf{I} - \mathbf{W}_h)\chi$, with $\mathbf{W}_h = (w_h(\chi_i,\chi_j))_{1\le i,j\le n-1}$, $\mathbf{X} = (x_{ij})_{1\le i\le n-1, 1\le j\le p}$ and $\chi(\bullet) = (\chi_2(\bullet),\ldots,\chi_n(\bullet))^T$, the estimator for β is defined by

$$\widehat{\beta}_h(\bullet) = (\widetilde{\mathbf{X}}_h^T \widetilde{\mathbf{X}}_h)^{-1} \widetilde{\mathbf{X}}_h^T \widetilde{\chi}_h(\bullet). \tag{2.5}$$