

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](http://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](mailto: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

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*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](http://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 account 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 brilliant 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), trachtophography 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 IWFO'S'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 IWFO'S'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 efficient 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, Australia), S. Marron (North Carolina, USA), P. Sarda (Toulouse, France), M. Valderama (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-



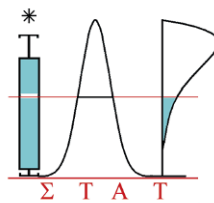
ysis and High-Dimensional Data contributed to the development of numerous scientific collaborations with statisticians in the whole world. A first consequence was the creation, the organization and the management of the first edition of IWFOs (Toulouse, France, 2008). The success of IWFOs’2008 was certainly 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

<b>Preface</b> .....	v
<b>Acknowledgements</b> .....	vii
<b>1 Penalized Spline Approaches for Functional Principal Component Logit Regression</b> .....	1
A. Aguilera, M. C. Aguilera-Morillo, M. Escabias, M. Valderrama	
1.1 Introduction .....	1
1.2 Background .....	2
1.3 Penalized estimation of FPCLR .....	3
1.3.1 Functional PCA via P-splines .....	4
1.3.2 P-spline smoothing of functional PCA .....	4
1.4 Simulation study .....	5
References .....	6
<b>2 Functional Prediction for the Residual Demand in Electricity Spot Markets</b> .....	9
Germán Aneiros, Ricardo Cao, Juan M. Vilar-Fernández, Antonio Muñoz-San-Roque	
2.1 Introduction .....	9
2.2 Functional nonparametric model .....	11
2.3 Semi-functional partial linear model .....	12
2.4 Data description and empirical study .....	13
References .....	14
<b>3 Variable Selection in Semi-Functional Regression Models</b> .....	17
Germán Aneiros, Frédéric Ferraty, Philippe Vieu	
3.1 Introduction .....	17
3.2 The methodology .....	18
3.3 Asymptotic results .....	20
3.4 A simulation study .....	20
References .....	22

<b>4</b>	<b>Power Analysis for Functional Change Point Detection</b> . . . . .	23
	John A. D. Aston, Claudia Kirch	
	4.1 Introduction . . . . .	23
	4.2 Testing for a change . . . . .	24
	4.3 Asymptotic Power Analysis . . . . .	25
	References . . . . .	26
<b>5</b>	<b>Robust Nonparametric Estimation for Functional Spatial Regression</b> . . . . .	27
	Mohammed K. Attouch, Abdelkader Gheriballah, Ali Laksaci	
	5.1 Introduction . . . . .	27
	5.2 The model . . . . .	28
	5.3 Main results . . . . .	29
	References . . . . .	31
<b>6</b>	<b>Sequential Stability Procedures for Functional Data Setups</b> . . . . .	33
	Alexander Aue, Siegfried Hörmann, Lajos Horváth, Marie Hušková	
	6.1 Introduction . . . . .	33
	6.2 Test procedures . . . . .	34
	6.3 Asymptotic properties . . . . .	37
	References . . . . .	38
<b>7</b>	<b>On the Effect of Noisy Observations of the Regressor in a Functional Linear Model</b> . . . . .	41
	Mareike Bereswill, Jan Johannes	
	7.1 Introduction . . . . .	41
	7.2 Background to the methodology . . . . .	43
	7.3 The effect of noisy observations of the regressor . . . . .	45
	References . . . . .	47
<b>8</b>	<b>Testing the Equality of Covariance Operators</b> . . . . .	49
	Graciela Boente, Daniela Rodriguez, Mariela Sued	
	8.1 Introduction . . . . .	49
	8.2 Notation and preliminaries . . . . .	50
	8.3 Hypothesis Test . . . . .	51
	8.4 Generalization to k-populations . . . . .	52
	References . . . . .	53
<b>9</b>	<b>Modeling and Forecasting Monotone Curves by FDA</b> . . . . .	55
	Paula R. Bouzas, Nuria Ruiz-Fuentes	
	9.1 Introduction . . . . .	55
	9.2 Functional reconstruction of monotone sample paths . . . . .	56
	9.3 Modeling and forecasting . . . . .	57
	9.4 Application to real data . . . . .	59
	9.5 Conclusions . . . . .	59
	References . . . . .	60

- 10 Wavelet-Based Minimum Contrast Estimation of Linear Gaussian Random Fields** . . . . . 63  
 Rosa M. Crujeiras, María-Dolores Ruiz-Medina
  - 10.1 Introduction . . . . . 63
  - 10.2 Wavelet generalized RFs . . . . . 64
  - 10.3 Consistency of the wavelet periodogram . . . . . 66
  - 10.4 Minimum contrast estimator . . . . . 68
  - 10.5 Final comments . . . . . 69
  - References . . . . . 69
  
- 11 Dimensionality Reduction for Samples of Bivariate Density Level Sets: an Application to Electoral Results** . . . . . 71  
 Pedro Delicado
  - 11.1 Introduction . . . . . 71
  - 11.2 Multidimensional Scaling for density level datasets . . . . . 73
  - 11.3 Analyzing electoral behavior . . . . . 74
  - References . . . . . 75
  
- 12 Structural Tests in Regression on Functional Variable** . . . . . 77  
 Laurent Delsol, Frédéric Ferraty, Philippe Vieu
  - 12.1 Introduction . . . . . 77
  - 12.2 Structural tests . . . . . 78
    - 12.2.1 A general way to construct a test statistic . . . . . 78
    - 12.2.2 Bootstrap methods to get the threshold . . . . . 80
  - 12.3 Application in spectrometry . . . . . 81
  - 12.4 Discussion and prospects . . . . . 81
  - References . . . . . 82
  
- 13 A Fast Functional Locally Modeled Conditional Density and Mode for Functional Time-Series** . . . . . 85  
 Jacques Demongeot, Ali Laksaci, Fethi Madani, Mustapha Rachdi
  - 13.1 Introduction . . . . . 85
  - 13.2 Main results . . . . . 86
  - 13.3 Interpretations and remarks . . . . . 88
  - References . . . . . 90
  
- 14 Generalized Additive Models for Functional Data** . . . . . 91  
 Manuel Febrero-Bande, Wenceslao González-Manteiga
  - 14.1 Introduction . . . . . 91
  - 14.2 Transformed Binary Response Regression Models . . . . . 92
  - 14.3 GAM: Estimation and Prediction . . . . . 93
  - 14.4 Application . . . . . 94
  - References . . . . . 96

<b>15</b>	<b>Recent Advances on Functional Additive Regression</b> . . . . .	97
	Frédéric Ferraty, Aldo Goia, Enersto Salinelli, Philippe Vieu	
15.1	The additive decomposition . . . . .	97
15.2	Construction of the estimates . . . . .	98
15.3	Theoretical results . . . . .	100
15.4	Application to real and simulated data . . . . .	101
	References . . . . .	102
<b>16</b>	<b>Thresholding in Nonparametric Functional Regression with Scalar Response</b> . . . . .	103
	Frédéric Ferraty, Adela Martínez-Calvo, Philippe Vieu	
16.1	Introduction . . . . .	103
16.2	Threshold estimator . . . . .	104
16.3	Cross-validation criterion: a graphical tool . . . . .	105
16.4	Simulation study . . . . .	107
	References . . . . .	109
<b>17</b>	<b>Estimation of a Functional Single Index Model</b> . . . . .	111
	Frédéric Ferraty, Juhyun Park, Philippe Vieu	
17.1	Introduction . . . . .	111
17.2	Index parameter as an <i>average</i> derivative . . . . .	113
17.3	Estimation of the directional derivatives . . . . .	114
17.4	Estimation for functional single index model . . . . .	115
	References . . . . .	115
<b>18</b>	<b>Density Estimation for Spatial-Temporal Data</b> . . . . .	117
	Liliana Forzani, Ricardo Fraiman, Pamela Llop	
18.1	Introduction . . . . .	117
18.2	Density estimator . . . . .	118
18.2.1	Stationary case: $\mu(\mathbf{s}) = \mu$ constant . . . . .	119
18.2.2	Non-stationary case: $\mu(\mathbf{s})$ any function . . . . .	119
18.2.3	Hypothesis . . . . .	120
18.2.4	Asymptotic results . . . . .	120
	References . . . . .	121
<b>19</b>	<b>Functional Quantiles</b> . . . . .	123
	Ricardo Fraiman, Beatriz Pateiro-López	
19.1	Introduction . . . . .	123
19.2	Quantiles in Hilbert spaces . . . . .	124
19.2.1	Sample quantiles . . . . .	125
19.2.2	Asymptotic behaviour . . . . .	126
19.3	Principal quantile directions . . . . .	127
19.3.1	Sample principal quantile directions . . . . .	128
19.3.2	Consistency of principal quantile directions . . . . .	128
	References . . . . .	129

**20 Extremality for Functional Data** . . . . . 131  
 Alba M. Franco-Pereira, Rosa E. Lillo, Juan Romo  
 20.1 Introduction . . . . . 131  
 20.2 Two measures of extremality for functional data . . . . . 132  
 20.3 Finite-dimensional versions . . . . . 133  
 References . . . . . 134

**21 Functional Kernel Estimators of Conditional Extreme Quantiles** . . . . 135  
 Laurent Gardes, Stéphane Girard  
 21.1 Introduction . . . . . 135  
 21.2 Notations and assumptions . . . . . 136  
 21.3 Main results . . . . . 137  
 References . . . . . 140

**22 A Nonparametric Functional Method for Signature Recognition** . . . . 141  
 Gery Geenens  
 22.1 Introduction . . . . . 141  
 22.2 Signatures as random objects . . . . . 142  
 22.3 A semi-normed functional space for signatures . . . . . 143  
 22.4 Nonparametric functional signature recognition . . . . . 144  
 22.5 Concluding remarks . . . . . 146  
 References . . . . . 147

**23 Longitudinal Functional Principal Component Analysis** . . . . . 149  
 Sonja Greven, Ciprian Crainiceanu, Brian Caffo, Daniel Reich  
 23.1 Introduction . . . . . 149  
 23.2 The Longitudinal Functional Model and LFPCA . . . . . 151  
 23.3 Estimation and Simulation Results . . . . . 152  
 23.4 Application to the Tractography Data . . . . . 153  
 References . . . . . 153

**24 Estimation and Testing for Geostatistical Functional Data** . . . . . 155  
 Oleksandr Gromenko, Piotr Kokoszka  
 24.1 Introduction . . . . . 155  
 24.2 Estimation of the mean function . . . . . 157  
 24.3 Estimation of the functional principal components . . . . . 159  
 24.4 Applications to inference for spatially distributed curves . . . . . 159  
 References . . . . . 160

**25 Structured Penalties for Generalized Functional Linear Models (GFLM)** . . . . . 161  
 Jaroslaw Harezlak, Timothy W. Randolph  
 25.1 Introduction . . . . . 161  
 25.2 Overview of PEER . . . . . 162  
     25.2.1 Structured and targeted penalties . . . . . 164  
     25.2.2 Analytical properties . . . . . 165  
 25.3 Extension to GFLM . . . . . 165

25.4	Application to a magnetic resonance spectroscopy data . . . . .	165
25.5	Discussion . . . . .	166
	References . . . . .	167
<b>26</b>	<b>Consistency of the Mean and the Principal Components of Spatially Distributed Functional Data</b> . . . . .	<b>169</b>
	Siegfried Hörmann, Piotr Kokoszka	
26.1	Introduction . . . . .	169
26.2	Model and dependence assumptions . . . . .	170
26.3	The sampling schemes . . . . .	172
26.4	Some selected results . . . . .	173
	References . . . . .	175
<b>27</b>	<b>Kernel Density Gradient Estimate</b> . . . . .	<b>177</b>
	Ivana Horová, Kamila Vopatová	
27.1	Kernel density estimator . . . . .	177
27.2	Kernel gradient estimator . . . . .	177
27.3	A proposed method . . . . .	179
27.4	Simulations . . . . .	181
	References . . . . .	182
<b>28</b>	<b>A Backward Generalization of PCA for Exploration and Feature Extraction of Manifold-Valued Shapes</b> . . . . .	<b>183</b>
	Sungkyu Jung	
28.1	Introduction . . . . .	183
28.2	Finite and infinite dimensional shape spaces . . . . .	184
28.3	Principal Nested Spheres . . . . .	185
28.4	Conclusion . . . . .	186
	References . . . . .	187
<b>29</b>	<b>Multiple Functional Regression with both Discrete and Continuous Covariates</b> . . . . .	<b>189</b>
	Hachem Kadri, Philippe Preux, Emmanuel Duflos, Stéphane Canu	
29.1	Introduction . . . . .	189
29.2	Multiple functional regression . . . . .	191
29.3	Conclusion . . . . .	194
	References . . . . .	194
<b>30</b>	<b>Combining Factor Models and Variable Selection in High-Dimensional Regression</b> . . . . .	<b>197</b>
	Alois Kneip, Pascal Sarda	
30.1	Introduction . . . . .	197
30.2	The augmented model . . . . .	199
30.3	Estimation . . . . .	200
30.4	Theoretical properties of augmented model . . . . .	201
	References . . . . .	202

**31 Factor Modeling for High Dimensional Time Series** . . . . . 203  
 Clifford Lam, Qiwei Yao, Neil Bathia  
 31.1 Introduction . . . . . 203  
 31.2 Estimation Given  $r$  . . . . . 205  
 31.3 Determining  $r$  . . . . . 206  
 References . . . . . 206

**32 Depth for Sparse Functional Data** . . . . . 209  
 Sara López-Pintado, Ying Wei  
 32.1 Introduction . . . . . 209  
 32.2 Method . . . . . 210  
     32.2.1 Review on band depth and modified band depth . . . . . 210  
     32.2.2 Adapted conditional depth for sparse data . . . . . 211  
 References . . . . . 212

**33 Sparse Functional Linear Regression with Applications to Personalized Medicine** . . . . . 213  
 Ian W. McKeague, Min Qian  
 33.1 Introduction . . . . . 213  
 33.2 Threshold-based point impact treatment policies . . . . . 214  
 33.3 Assessing the estimated TPI policy . . . . . 216  
 References . . . . . 217

**34 Estimation of Functional Coefficients in Partial Differential Equations** . . . . . 219  
 Jose C. S. de Miranda  
 34.1 Introduction . . . . . 219  
 34.2 Estimator construction . . . . . 220  
 34.3 Main results . . . . . 222  
 34.4 Final remarks . . . . . 223  
 References . . . . . 224

**35 Functional Varying Coefficient Models** . . . . . 225  
 Hans-Georg Müller, Damla Şentürk  
 35.1 Introduction . . . . . 225  
 35.2 Varying coefficient models with history index . . . . . 226  
 35.3 Functional approach for the ordinary varying coefficient model . . . . . 227  
 35.4 Fitting the history index model . . . . . 228  
 References . . . . . 230

**36 Applications of Funtional Data Analysis to Material Science** . . . . . 231  
 S. Naya, M. Francisco-Fernández, J. Tarrío-Saavedra, J. López-Beceiro, R. Artiaga  
 36.1 Introduction . . . . . 231  
 36.2 Materials testing and data collecting . . . . . 232  
 36.3 Statistical methods . . . . . 233  
 36.4 Results and discussion . . . . . 234



36.5 New research lines ..... 236  
 References ..... 237

**37 On the Properties of Functional Depth** ..... 239  
 Alicia Nieto-Reyes

37.1 Introduction ..... 239  
 37.2 Properties of functional depth ..... 240  
 37.3 A well-behave functional depth ..... 242  
 37.4 Conclusions ..... 243  
 References ..... 243

**38 Second-Order Inference for Functional Data with Application to DNA Minicircles** ..... 245  
 Victor M. Panaretos, David Kraus, John H. Maddocks

38.1 Introduction ..... 245  
 38.2 Test ..... 247  
 38.3 Application to DNA minicircles ..... 249  
 References ..... 250

**39 Nonparametric Functional Time Series Prediction** ..... 251  
 Efsthios Paparoditis

39.1 Wavelet-kernel based prediction ..... 251  
 39.2 Bandwidth Choice ..... 253  
 39.3 Further Issues ..... 254  
 References ..... 254

**40 Wavelets Smoothing for Multidimensional Curves** ..... 255  
 Davide Pigoli, Laura M. Sangalli

40.1 Introduction ..... 255  
 40.2 An overview on wavelets ..... 256  
 40.3 Wavelet estimation for  $p$  - dimensional curves ..... 257  
 40.4 Application to ECG data ..... 259  
 References ..... 260

**41 Nonparametric Conditional Density Estimation for Functional Data. Econometric Applications** ..... 263  
 Alejandro Quintela-del-Río, Frédéric Ferraty, Philippe Vieu

41.1 Introduction ..... 263  
 41.2 The conditional density estimator ..... 264  
 41.3 Testing a parametric form for the conditional density ..... 264  
 41.4 Value-at-risk and expected shortfall estimation ..... 265  
 41.5 Simulations ..... 266  
     41.5.1 Results for the hypothesis testing. .... 266  
     41.5.2 Results for the CVaR and CES estimates ..... 267  
 References ..... 268

**42 Spatial Functional Data Analysis** . . . . . 269  
 James O. Ramsay, Tim Ramsay, Laura M. Sangalli

42.1 Introduction . . . . . 269  
 42.2 Data, model and estimation problem . . . . . 270  
 42.3 Finite element solution of the estimation problem . . . . . 272  
 42.4 Simulations . . . . . 272  
 42.5 Discussion . . . . . 273  
 References . . . . . 275

**43 Clustering Spatially Correlated Functional Data** . . . . . 277  
 Elvira Romano, Ramon Giraldo, Jorge Mateu

43.1 Introduction . . . . . 277  
 43.2 Spatially correlated functional data . . . . . 278  
 43.3 Hierarchical clustering of spatially correlated functional data . . . . . 279  
 43.4 Dynamic clustering of spatially correlated functional data . . . . . 280  
 43.5 Discussion . . . . . 281  
 References . . . . . 282

**44 Spatial Clustering of Functional Data** . . . . . 283  
 Piercesare Secchi, Simone Vantini, Valeria Vitelli

44.1 Introduction . . . . . 283  
 44.2 A clustering procedure for spatially dependent functional data . . . . . 284  
 44.3 A simulation study on synthetic data . . . . . 285  
 44.4 A case study: clustering irradiance data . . . . . 287  
 References . . . . . 289

**45 Population-Wide Model-Free Quantification of Blood-Brain-Barrier Dynamics in Multiple Sclerosis** . . . . . 291  
 Russell Shinohara, Ciprian Crainiceanu

45.1 Introduction . . . . . 291  
 45.2 Methods and Results . . . . . 292  
 45.3 Conclusions . . . . . 295  
 References . . . . . 296

**46 Flexible Modelling of Functional Data using Continuous Wavelet Dictionaries** . . . . . 297  
 Leen Slaets, Gerda Claeskens, Maarten Jansen

46.1 Introduction . . . . . 297  
 46.2 Modelling Functional Data by means of Continuous Wavelet dictionaries . . . . . 298  
 References . . . . . 300

**47 Periodically Correlated Autoregressive Hilbertian Processes of Order  $p$**  . . . . . 301  
 Ahmad R. Soltani, Majid Hashemi

47.1 Introduction . . . . . 301  
 47.2 Large Sample Theorems . . . . . 304

- 47.3 Parameter estimation ..... 305
- References ..... 306
  
- 48 Bases Giving Distances. A New Semimetric and its Use for  
Nonparametric Functional Data Analysis ..... 307**  
Catherine Timmermans, Laurent Delsol, Rainer von Sachs
- 48.1 Introduction ..... 307
- 48.2 Definition of the semimetric ..... 308
- 48.3 Nonparametric functional data analysis ..... 310
- References ..... 313
  
- List of Contributors ..... 315**

# 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](mailto:aaguiler@ugr.es)

Maria del Carmen Aguilera-Morillo

Department of Statistics and O. R. University of Granada, Spain, e-mail: [caguilera@ugr.es](mailto:caguilera@ugr.es)

Manuel Escabias

Department of Statistics and O. R. University of Granada, Spain, e-mail: [escabias@ugr.es](mailto:escabias@ugr.es)

Mariano Valderrama

Department of Statistics and O. R. University of Granada, Spain, e-mail: [valderra@ugr.es](mailto:valderra@ugr.es)

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

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

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

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

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

The logit transformations can be expressed as

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

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

$$L = \alpha \mathbf{1} + \Gamma \gamma, \quad (1.2)$$

where  $\Gamma = (\xi_{ij})_{n \times p}$  is a matrix of functional pc's of  $x_1(t), \dots, x_n(t)$  and  $\gamma$  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  $\beta$  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  $i^{\text{th}}$  observed value of the binary response and  $\hat{y}_i^{(-i)}$  the  $i^{\text{th}}$  predicted value obtained by deleting the  $i^{\text{th}}$  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  $\langle f, g \rangle = \int_T f(t)g(t)dt$ . In the standard formulation of functional PCA, the  $j^{\text{th}}$  principal component scores are given by

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

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

$$\begin{cases} \text{Max}_f \text{Var}[\int_T x_i(t) f(t) dt] \\ \text{s.t. } \|f\|^2 = 1 \text{ and } \int f_\ell(t) f(t) dt = 0, \quad \ell = 1, \dots, j-1. \end{cases}$$

The weight functions  $f_j$  are the solutions to the eigenequation  $Cf_j = \lambda_j f_j$ , with  $\lambda_j = \text{Var}[\xi_j]$  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  $\{t_{i0}, t_{i1}, \dots, t_{im_i} \in T, i = 1, \dots, n\}$ . The sample information is given by the vectors  $x_i = (x_{i0}, \dots, x_{im_i})'$ , with  $x_{ik}$  the observed value for the  $i^{\text{th}}$  sample path  $x_i(t)$  at time  $t_{ik}$  ( $k = 0, \dots, m_i$ ).

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  $a_i = (a_{i1}, \dots, a_{ip})'$ .

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  $P_d = (\Delta^d)' \Delta^d$  and  $\Delta^d$  is the differencing matrix that gives the  $d$ th-order differences of  $a_i$ . The solution is then given by  $\hat{a}_i = (\Phi_i' \Phi_i + \lambda P_d)^{-1} \Phi_i' x_i$ , 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  $j^{\text{th}}$  principal component scores are defined as in equation (1.3) but now the weight functions  $f_j$  are obtained by solving

$$\begin{cases} \text{Max}_f \frac{\text{var}[\int x_i(t) f(t) dt]}{\|f\|^2 + \lambda \text{PEN}_d(f)} \\ \text{s.t. } \|f\|^2 = b' \Psi b = 1 \text{ and } b' \Psi b_\ell + b' P_d b_\ell = 0, \ell = 1, \dots, j-1, \end{cases}$$

where  $\text{PEN}_d(f) = b' P_d b$  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  $\lambda$  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  $\beta$  of basis coefficients of the functional parameter is given by  $\hat{\beta} = F\hat{\gamma} = (L^{-1})'G\hat{\gamma}$ , 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.

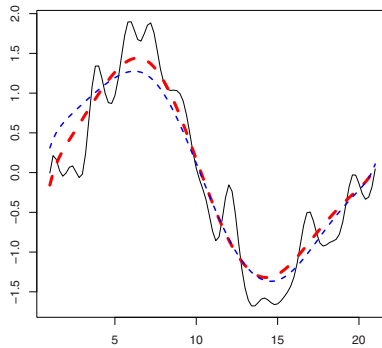


Fig. 1.1: Estimated parameter function with the three different considered FPCA estimations: non smoothed FPCA (black and continue line), FPCA via P-splines (red and long dashed line,  $\lambda = 24.2$ ) and P-spline smoothed FPCA (blue and short dashed line,  $\lambda = 5$ )

**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](mailto:ganeiros@udc.es)

Ricardo Cao  
Universidade da Coruña, Spain, e-mail: [rcao@udc.es](mailto:rcao@udc.es)

Juan M. Vilar-Fernández  
Universidade da Coruña, Spain, e-mail: [eijvilar@udc.es](mailto:eijvilar@udc.es)

Antonio Muñoz-San-Roque  
Universidad Pontificia de Comillas, Madrid, Spain, e-mail: [antonio.munoz@iit.icaui.upcomillas.es](mailto:antonio.munoz@iit.icaui.upcomillas.es)

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 Gooijer (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,  $r$ , ( $r \in \{1, \dots, 24\}$ ) 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  $\{\chi_i^{(r)}(p)\}_{i=1,2,\dots,n}$ . 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

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

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  $\mathcal{H} = \{f : C \rightarrow \mathbb{R}\}$  the space where the functional data live and by  $d(\bullet, \bullet)$  a semi-metric associated with  $\mathcal{H}$ . Thus  $(\mathcal{H}, d)$  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

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

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

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

and the kernel function  $K : [0, \infty) \rightarrow [0, \infty)$  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 sparseness 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:

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

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

Now, based on the SFPL model, we may look at the problem of predicting  $\chi_{n+1}(\bullet)$  by computing estimations  $\hat{\beta}$  and  $\hat{m}(\chi)$  of  $\beta$  and  $m(\chi)$  in (2.4), respectively. Thus,  $\mathbf{x}_{n+1}^T \hat{\beta}(\bullet) + \hat{m}(\chi_n)$  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  $\tilde{\mathbf{X}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{X}$  and  $\tilde{\chi}_h = (\mathbf{I} - \mathbf{W}_h)\chi$ , with  $\mathbf{W}_h = (w_h(\chi_i, \chi_j))_{1 \leq i, j \leq n-1}$ ,  $\mathbf{X} = (x_{ij})_{1 \leq i \leq n-1, 1 \leq j \leq p}$  and  $\chi(\bullet) = (\chi_2(\bullet), \dots, \chi_n(\bullet))^T$ , the estimator for  $\beta$  is defined by

$$\hat{\beta}_h(\bullet) = (\tilde{\mathbf{X}}_h^T \tilde{\mathbf{X}}_h)^{-1} \tilde{\mathbf{X}}_h^T \tilde{\chi}_h(\bullet). \quad (2.5)$$