

UNIVERSITY OF GRANADA



**PhD PROGRAM IN
MATHEMATICS AND STATISTICS
(D05.56.1)**

**PENALIZED ESTIMATION METHODS
IN FUNCTIONAL DATA ANALYSIS**

PhD DISSERTATION

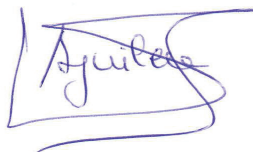
**María del Carmen Aguilera Morillo
Granada, April, 2013**

Editor: Editorial de la Universidad de Granada
Autor: María del Carmen Aguilera Morillo
D.L.: GR 2244-2013
ISBN: 978-84-9028-649-4

La doctoranda María del Carmen Aguilera Morillo y la directora de la tesis Dña. Ana María Aguilera del Pino. Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección de la directora de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados cuando se han utilizado sus resultados o publicaciones.

Granada, Abril de 2013.

Directora de la Tesis

A handwritten signature in blue ink, appearing to read 'Aguilera', enclosed within a rectangular box drawn with the same ink.

Fdo.: Dña. Ana M. Aguilera del Pino.

Doctoranda

A handwritten signature in blue ink, appearing to read 'M. Carmen A.', enclosed within a circular shape drawn with the same ink.

Fdo.: Dña. M. Carmen Aguilera Morillo.

A Carlos

Contents

	Page
Introduction	1
1 Smoothing with B-spline bases	13
1.1 Introduction	13
1.2 Basic tools for FDA	15
1.3 Basis expansion of functional data	18
1.3.1 B-splines	19
1.3.2 Other bases	21
1.4 Smoothing with B-spline bases	23
1.4.1 Regression splines	23
1.4.2 Smoothing splines	25
1.4.3 P-splines	26
1.5 Choosing the smoothing parameter	28
1.6 Simulation study	30
1.7 Real data applications	35
1.7.1 Pinch data	35
1.7.2 Manure data	37
1.8 Conclusions	39
2 Penalized PCA approaches for B-spline expansions of smooth functional data	41
2.1 Introduction	41
2.2 Functional principal component analysis	44
2.2.1 Basis expansion estimation	46
2.3 P-spline smoothed functional PCA	47
2.3.1 Selection of the smoothing parameter	49
2.4 Functional PCA of P-splines	50

2.4.1	Selection of the smoothing parameter	51
2.5	Simulation study	52
2.6	Real data application	58
2.7	Computational cost	60
2.8	Conclusions	63
3	Penalized spline approaches for functional logit regression	65
3.1	Introduction	65
3.2	Functional logit model	68
3.2.1	Penalized estimation with basis expansions	70
3.3	Penalized estimation of functional principal component logit regression	71
3.3.1	Method I: non-penalized FPCLoR	72
3.3.2	Method II: FPCLoR on P-spline smoothing of the sample curves	73
3.3.3	Method III: FPCLoR on P-spline smoothing of the principal components	73
3.3.4	Method IV: FPCLoR with P-spline penalty in the maximum likelihood estimation	74
3.4	Model Selection	75
3.4.1	Choosing λ in Method II	75
3.4.2	Choosing λ in Method III	76
3.4.3	Choosing the number of principal components in Methods I, II, and III	76
3.4.4	Choosing the number of predictors and the smoothing parameter in Methods IV and V	77
3.5	Simulation study	78
3.5.1	Case I: simulation of waveform data	78
3.5.2	Case II: simulation of the Ornstein-Uhlenbeck process	83
3.6	Conclusions	85
4	Penalized spline approaches for functional PLS regression	91
4.1	Introduction	91
4.2	Functional PLS	93
4.2.1	Basis expansion estimation	96
4.3	Penalized functional PLS	98
4.3.1	Roughness penalty function	98
4.3.2	FPLS by penalizing the norm	99

4.3.3	FPLS by penalizing the covariance	101
4.3.4	Sample estimation	102
4.3.5	Model selection	104
4.4	Simulation study	105
4.4.1	Description	105
4.4.2	Discussion of results	108
4.5	Real data application	115
4.6	Conclusions	117
5	P-spline estimation of functional classification methods for improving the quality in the food industry	121
5.1	Introduction	121
5.2	Smoothing the data	124
5.3	Methodological aspects	126
5.3.1	Functional principal component logit regression	128
5.3.2	Functional linear discriminant analysis based on functional PLS regression	131
5.3.3	Componentwise classification	132
5.4	Results	134
5.4.1	Interpreting the weight function	136
5.5	Conclusions	141
A	Software and computational considerations	145
A.1	Main libraries	146
A.2	Main functions	146
A.2.1	Smoothing with B-spline bases	147
A.2.2	Functional PCA	148
A.2.3	Penalized functional PC logit regression	149
A.2.4	Penalized functional PLS regression	149
B	Conclusions and further research	151
B.1	Chapter 1	151
B.2	Chapter 2	152
B.3	Chapter 3	153
B.4	Chapter 4	155
B.5	Chapter 5	156
B.6	Further research	157

C Conclusiones y líneas abiertas	159
C.1 Capítulo 1	159
C.2 Capítulo 2	160
C.3 Capítulo 3	161
C.4 Capítulo 4	163
C.5 Capítulo 5	165
C.6 Líneas abiertas	166
D Summary	167
D.1 Chapter 1	170
D.2 Chapter 2	171
D.3 Chapter 3	173
D.4 Chapter 4	176
D.5 Chapter 5	179
E Resumen	183
E.1 Capítulo 1	186
E.2 Capítulo 2	187
E.3 Capítulo 3	190
E.4 Capítulo 4	193
E.5 Capítulo 5	196
Bibliography	199
List of Figures	211
List of Tables	215

Introduction

A functional variable is characterized because its observations are functions that in the majority of cases represent the evolution of a scalar variable in time (realizations of a stochastic process). This is the case of environmental variables such as temperature or contamination level observed daily in a period of time, economic variables such as stock price evolution or medical variables such as stress level. In other areas of application the argument of the observed functions is a different magnitude such as spatial location, wavelength or probability. In many chemometric applications, observations of the NIR spectrum at a fine grid of wavelengths are available.

Functional data analysis (FDA) is an statistical topic of active research devoted to solve problems related with the statistical modeling and prediction of functional data. An overview of the basic methods of FDA, computational aspects related with their practical application and important real data modeling can be seen in the pioneers books by Ramsay and Silverman (1997, 2005, 2002) and Ramsay et al. (2009). A detailed study on nonparametric FDA methodologies was developed in Ferraty and Vieu (2006). Statistical inference related with some FDA methods was recently studied in Horvath and Kokoszka (2012).

The interest of FDA is reflected in the growing number of articles on this question in recent years, from the two papers that appeared in 1997 to the 83 published in 2011, according to the ISI Web of Knowledge database. This growth became particularly evident following the publication of the first specialized book in the field (Ramsay and Silverman, 1997). From a practical point of view, different fields where this topic has aroused special interest are health sciences and biology, where in recent years 222 articles have been published. A revision of the FDA methods usually used in Biometrics and Biostatistics and interesting applications can be seen in Escabias et al. (2012).

Early work on FDA was developed in the framework of continuous-time stochastic processes and was devoted to the generalization of reduction dimension techniques such as principal component analysis (PCA) to the functional case (Deville, 1974). Later, statistical researching on FDA focused on the formulation and estimation of different functional regression models. The functional linear model to estimate a scalar response variable from a functional predictor was one of the first regression models extended to the functional data case (Cardot et al., 1999, 2003). The case where the predictor is a vector or scalar and the response is functional was studied by Chiou et al. (2004). Functional analysis of variance was introduced to model the mean of a functional response in terms of a categorical variable (Cuevas et al., 2002, 2004). On the other hand, functional linear models where both predictor and response variables are functional were studied by Yao et al. (2005b) and Ocaña et al. (2008). Principal component prediction models, that can be seen as a particular case of these linear models, were first introduced to forecast a continuous time stochastic process on a future interval from its recent past (Aguilera et al., 1997, 1999). Generalized linear models were also extended to the case of a functional predictor (James, 2002; Müller, 2005). A particular case of functional generalized model is the functional logit regression model whose aim is to predict a binary random variable from a functional predictor (Ratcliffe et al., 2002; Escabias et al., 2004; Aguilera et al., 2008b). On the other hand, a spatial spline regression model for the analysis of data distributed over irregularly shaped spatial domains is proposed in Sangalli et al. (2013).

Direct estimation of the functional parameter associated with a functional regression model is an ill-posed problem due to the infinite dimension of the functional variable. On the other hand, sample curves are usually observed in a finite set of sampling points that could be unequally spaced and different among the sample units. Because of this, the first step in FDA is to reconstruct the true functional form of each sample curve from a finite set of discrete observations. Approximation techniques such as interpolation or projection in a finite-dimensional space generated by basis functions were applied from the beginning to solve these problems. This way, the estimation of a functional regression model is reduced to the estimation of an equivalent multivariate regression model with high correlation between the predictor variables.

Regression on a set of uncorrelated random variables is usually used in

literature to provide an accurate estimation of the parameters associated with a regression model. Functional PCA was used to reduce the dimension and solving the multicollinearity problem in many functional regression models. Principal components are uncorrelated generalized linear combinations of the functional predictor with maximum variance. Because of this the main criticism about principal component regression is that the regressors are computed without taking into account the response variable. To solve this problem, functional partial least squares (PLS) was extended to the functional case by computing a set of uncorrelated generalized linear combinations of the predictor variable having maximum covariance with the response variable (Preda and Saporta, 2005b).

In many applications the data are smooth functions observed with error. In this case least squares approximation with B-spline bases is usually used to estimate the basis coefficients of a basis expansion of the unobserved smooth sample functions. The problem is that the approximated sample curves (regression splines) do not control the degree of smoothness. As a consequence, the estimated principal components and functional parameters associated with functional regression models are difficult to interpret because they have a lot of variability and lack of smoothness.

The general objective of this thesis is to improve the estimation of FDA methodologies in the case of smooth functional data observed with error. In order to solve this problem, different approaches based on penalized estimation with B-spline basis expansions of sample curves are proposed. This general objective is achieved through five specific objectives:

1. Review and comparison of existing methods for the approximation of smooth curves with B-splines bases.
2. Improve the estimation of functional PCA by introducing different penalized spline approaches.
3. Develop different penalized approaches for estimating the functional logit model based on penalized spline estimation of functional PCA.
4. Propose different penalized estimation approaches in functional PLS regression.
5. Develop an application of the proposed penalized estimation methodologies to improve the quality in food industry.

According with the specific objectives, the thesis is divided into five chapters with the methodology and results related with each one. The contents of each chapter have been included in different research papers actually submitted or accepted for publication in different JCR journals. In addition to the methodological contributions in each chapter, the proposed penalized FDA methods were applied on simulated and real data by developing own code with the free statistical software R (<http://www.r-project.org>). A brief description of the main libraries and functions used in this thesis can be seen in Appendix A at the end of this memory.

Chapter 1

The main purpose of this chapter is to review and compare three different approaches for approximating smooth sample curves observed with error in terms of B-spline basis: regression splines (non-penalized least squares approximation), smoothing splines (continuous roughness penalty based on the integrated squared d -order derivative of each sample curve) and P-splines (discrete roughness penalty based on d -order differences between coefficients of adjacent B-splines). The performance of these spline smoothing approaches is studied via a simulation study and several applications with real data. Cross validation and generalize cross validation are adapted to select a common smoothing parameter for all sample curves with the roughness penalty approaches.

The approximation of smooth noisy functions with B-spline bases is used in the estimation of a wide variety of FDA methodologies. This justifies the importance of a comparison among the main smoothing approaches in terms of B-splines, drawing conclusions that allow the researchers and practitioners to use the most powerful tool in each case.

The main results of this chapter are submitted for publication in the following paper (actually revised and resubmitted according to the reviewers comments, and waiting for the editor decision):

- Comparative study of different B-spline approaches for functional data
Authors: Aguilera, A. M. and Aguilera-Morillo, M. C.
Ref.: Mathematical and Computer Modelling, 2012, under revision (Aguilera and Aguilera-Morillo, 2012)

A part of this study was presented in the congress

- XXXII Congreso Nacional de Estadística e I.O. y VI Jornadas de Estadística Pública

Mode of participation: Oral contributed paper

Title: Técnicas de suavizado vía splines en análisis de datos funcionales

Authors: Aguilera, A. M. and Aguilera-Morillo, M. C.

Ref.: Libro de Actas XXXII Congreso Nacional de Estadística e I.O. y VI Jornadas de Estadística Pública (J. Costa Bouzas, R. Fernández Casal, M.A. Presedo Quindimil and J.M. Vilar Fernández, eds.), Orbigraf, 2010, p. 109-110

Organizer: Universidad da Coruña y SEIO

Celebration: A Coruña (España), 2010, 14-9/17-9

Chapter 2

Functional principal component analysis (FPCA) is a dimension reduction technique that explains the dependence structure of a functional data set in terms of uncorrelated variables. In many applications data are a set of smooth functions observed with error. In these cases the principal components are difficult to interpret because the estimated weight functions have a lot of variability and lack of smoothness. The most common way to solve this problem is based on penalizing the roughness of a function by its integrated squared d -order derivative.

In this chapter, two alternative forms of penalized FPCA based on B-spline basis expansions of sample curves and a P-spline penalty are proposed. The main difference between both smoothed FPCA approaches is that the first one uses the P-spline penalty in the least squares approximation of the sample curves in terms of a B-spline basis meanwhile the second one introduces the P-spline penalty in the orthonormality constraint of the algorithm that computes the principal components. Leave-one-out cross validation is adapted to select the smoothing parameter for these two smoothed FPCA approaches. A simulation study and an application with chemometric functional data are developed to test the performance of the proposed penalized approaches and to compare the results with non-penalized FPCA and regularized FPCA.

The main results of this chapter are included in the following paper

- Penalized PCA approaches for B-spline expansions of smooth functional data

Authors: Aguilera, A. M. and Aguilera-Morillo, M. C.

Ref.: Applied Mathematics and Computation, 2013, in press (DOI 10.1016/j.amc.2013.02.009) (Aguilera and Aguilera-Morillo, 2013)

Part of this study was presented in the congresses

- 19th Conference IASC-ERS, COMPSTAT'2010

Title: Different P-spline approaches for smoothed functional principal component analysis

Mode of participation: Oral contributed paper

Authors: Aguilera, A. M., Aguilera-Morillo, M. C., Escabias, M. and Valderrama, M. J.

Ref.: Proceedings in Computational Statistics 2010 (Y. Lechevallier and G. Saporta, eds.), Springer-Verlag, 2010, 641-648 (Aguilera et al., 2010a)

Organizer: CNAM, INRIA and International Association for Statistical Computing

Celebration: Paris (France), 2010 (22-8/27-8)

- I Reunión de Trabajo del Grupo Análisis de Datos Funcionales de la Sociedad Española de Estadística e I.O.

Mode of participation: Poster

Title: Suavización P-spline del análisis en componentes principales funcional

Authors: Aguilera-Morillo, M. C. and Aguilera, A. M.

Organizer: Grupo Análisis de Datos Funcionales de la SEIO

Celebration: Santander (España), 15-6-2011

Chapter 3

This chapter is devoted to improve the estimation of the functional logit model. The problem of multicollinearity associated with the estimation of this model can be solved by using a set of functional principal components as predictor variables. The functional parameter estimated by functional principal component logit regression is often non-smooth and then difficult to interpret. To solve this problem different penalized spline estimations of the functional logit model are proposed in this chapter. All of them are based on smoothed functional PCA and/or a discrete P-spline penalty in the log-likelihood criterion in terms of B-spline expansions of the sample curves and the functional parameter.

In the context of functional principal component logit regression, three different versions of penalized estimation approaches based on smoothed FPCA are introduced. On the one hand, FPCA of P-spline approximation of sample curves (Method II) is performed. On the other hand, a discrete P-spline penalty, that penalizes the roughness of the principal component weight functions, is included in the own formulation of FPCA (Method III). The third smoothed approach is carried out by introducing the penalty in the likelihood estimation of the functional parameter in terms of a reduced set of functional principal components (Method IV). Moreover, direct P-spline likelihood estimation in terms of B-spline functions is also considered (Method V). The ability of these smoothing approaches to provide an accurate estimation of the functional parameter and their classification performance with respect to non-penalized functional PCA are evaluated via simulation and application to real data. Leave-one-out cross validation and generalized cross validation are adapted to select the smoothing parameter and the number of principal components or basis functions associated with the considered approaches.

Part of the results of this chapter are published in the paper

- Penalized spline approaches for functional logit regression

Authors: Aguilera-Morillo, M. C., Aguilera, A. M., Escabias, M. and Valderrama, M. J.

Ref.: Test, 2012, in press (DOI 10.1007/s11749-012-0307-1) (Aguilera-Morillo et al., 2012)

The computational algorithm developed in this chapter for the estima-

tion of the functional logit approach without using principal components was recently applied to predict the probability of high levels of airborne pollen in terms of the best subset of related functional climatic variables in the following paper:

- Stepwise selection of functional covariates in forecasting peak levels of olive pollen
Authors: Escabias, M., Valderrama, M. J., Aguilera, A. M., Santofimia, M. E. and Aguilera-Morillo, M. C.
Ref.: Stochastic Environmental Research and Risk Assessment, 2013, 27, 367-376 (Escabias et al., 2013)

The contributions of this chapter were partially presented in the congress

- International Workshop on Functional and Operatorial Statistics (IW-FOS 2011)
Mode of participation: Oral contributed paper
Title: Penalized spline approaches for functional principal component logit regression
Authors: Aguilera, A. M., Aguilera-Morillo, M. C., Escabias, M. and Valderrama, M. J.
Ref.: Recent Advances in Functional Data Analysis and Related Topics (Collection Contributions to Statistics) (F. Ferraty, ed.), Physica Verlag, 2011, 1-7 (Aguilera et al., 2011)
Organizer: Universidad de Cantabria
Celebration: Santander (España), 2011 (16-6/18-6)

Chapter 4

The main problems associated with the functional linear model for a scalar response in terms of smooth curves observed with error are high dimension, multicollinearity and non-smooth estimation of the functional parameter. In order to solve the three problems at the same time, two different penalized approaches based on partial least squares regression are developed. The main difference between the two proposed approaches is the way in which

the penalty is introduced. The first approach introduces the penalty in the definition of the norm of the PLS component weight functions (Method II). The second one considers a penalized estimation of the covariance between the response and the PLS components (Method III). Discrete and continuous penalties are considered in terms of basis expansions of the sample curves. The selection of the optimum number of PLS components and the smoothing parameter is carried out by different criteria based on GCV errors and the integrated mean squared errors of the parameter function.

In order to test the performance of the proposed penalized FPLS approaches and to compare the results with non-penalized FPLS, a simulation study and an application with chemometric functional data are developed.

The results of this chapter will be submitted for publication in an appropriate JCR journal as soon as possible.

Part of the results of this chapter were presented in the following congresses:

- VII Colloquium Chemometricum Mediterraneum (CCM VII)
Mode of participation: Poster
Title: Functional analysis of chemometric data
Authors: Aguilera, A.M., Escabias, M., Valderrama, M. J. and Aguilera-Morillo, M.C.
Ref.: e-Proceedings CCM VII (ISBN: 978-84-937483-4-0; M.G. Bagur González, A. González Casado y N. Navas Iglesias, eds.), Ayuda a la Enseñanza, S.L., 2010, P01-48, 3 pp.
Organizer: Universidad de Granada
Celebration: Granada (España), 2010 (21-6/24-6)
- 5th International Conference of the ERCIM Working Group on Computing and Statistics (ERCIM'12)
Authors: Aguilera, A.M., Aguilera-Morillo, M.C. and Preda, C.
Mode of participation: Invited session
Title: Introducing a roughness penalty in the functional PLS regression model
Ref.: Book of Abstracts of ERCIM'11, p. 35

Organizer: University of Oviedo

Celebration: Oviedo (España), 2012 (1-12/3-12)

Chapter 5

The aim of this chapter is to improve the quality of cookies production by classifying them as good or bad from the curves of resistance of dough observed during the kneading process. As the predictor variable is functional, functional classification methodologies such as functional logit regression and functional linear discriminant analysis are considered. A P-spline approximation of the sample curves is proposed to improve the classification ability of these models and to suitably estimate the relationship between the quality of cookies and the resistance of dough. Inference results on the functional parameters and related odds ratios are obtained using the asymptotic normality of the maximum likelihood estimators under the classical regularity conditions. Finally, the classification results are compared with alternative functional data analysis approaches such as componentwise classification on the logit regression model.

The main results of this chapter are submitted for publication (actually revised and resubmitted according to the reviewers comments, and waiting for the editor decision) in the paper

- P-spline estimation of functional classification methods for improving the quality in the food industry

Authors: Aguilera-Morillo, M. C. and Aguilera, A. M.

Ref.: Communications in Statistics - Simulation and Computation, under revision (Aguilera-Morillo and Aguilera, 2012)

Part of these results were presented in the congresses

- XXXIII Congreso Nacional de Estadística e I.O. y VII Jornadas de Estadística Pública

Mode of participation: Oral contributed paper

Title: Modelos funcionales penalizados para la mejora de la calidad en la producción de la industria alimenticia

Authors: Aguilera-Morillo, M. C., Aguilera, A. M. and Valderrama, M. J.

Ref.: Libro de Actas XXXIII Congreso Nacional de Estadística e I.O. y VII Jornadas de Estadística Pública (Dpto. Estadística e I.O. Universidad Rey Juan Carlos), AFANIAS, 2012, p. 44

Organizador: Universidad Rey Juan Carlos y SEIO

Celebration: Madrid (España), 2012, 17-4/20-4

- Joint Meeting of y-BIS International Young Business and Industrial Statisticians and jSPE Young Portuguese Statisticians

Mode of participation: Invited session

Title: P-spline estimation of functional classification methods for improving the quality in the food industry

Authors: Aguilera, A. M. and Aguilera-Morillo, M. C.

Celebration: Lisbon (Portugal), 2012, (23-07/26-07)

Acknowledgements in Spanish

Esta investigación ha sido subvencionada por los proyectos P11-FQM-8068 de la *Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía* y MTM2010-20502 de la *Dirección General de Investigación del Ministerio de Educación y Ciencia de España*, y la Beca FPU de la Universidad de Granada concedida a la doctoranda para la realización de esta Tesis Doctoral bajo la dirección de la Profesora Ana María Aguilera. A la autora le gustaría también agradecer a Caroline Lévéder por proporcionarle los datos de Danone usados en el desarrollo de la aplicación del Capítulo 5.

A la autora le gustaría dar las gracias a la directora de esta memoria, la profesora Dña. Ana M^a Aguilera del Pino, por su entrega incondicional durante estos cuatro años de investigación, por inculcarme el valor y la satisfacción del trabajo bien hecho, por animarme en los momentos más difíciles de esta tesis y por ser mi Maestra, y ante todo mi amiga. Al profesor D. Mariano Valderrama Bonnet, quien ha apostado y apuesta por mi carrera universitaria, animándome y ayudándome cada día a ser mejor profesional y mejor persona. A ambos, por fomentar mi formación haciéndome partícipe de numerosos congresos de gran relevancia en el campo de la Estadística, así como de varias estancias de investigación cofinanciadas por los proyectos que

coordinan. A todos los miembros de mi grupo de investigación, por vuestras continuas muestras de apoyo.

Al profesor D. Cristian Preda de la Ecole Polytechnique Universitaire de la Université de Lille por su aportación en el desarrollo del Capítulo 4 de esta tesis, y por su hospitalidad durante los cuatro meses de mi estancia en Lille.

A D. Óscar Duro, gerente de Servicios de Análisis en Axesor, por permitirme compatibilizar el mundo de la empresa y la universidad, sin olvidar a mis compañeras de Axesor, quienes me han animado en los días más oscuros.

De un modo muy especial, agradecer a mis maestros del colegio de El Palomar su interés en mi educación. Agradecer a muchos de mis profesores de instituto y universidad, quienes me enseñaron que en el conocimiento está la libertad, y quienes me animaron a emprender este camino que espero y deseo poder continuar. Todos ellos son mi referente.

A mis amigos, por entender mis interminables días de estudio y mi ausencia en muchas de nuestras quedadas.

Por último y no por ello menos importante, quiero agradecer a mi familia directa y política su interés, y ante todo agradecer la educación recibida de mi madre, sin cuya ayuda y grandes esfuerzos hoy no estaría firmando esta tesis. Y no podría terminar sin agradecer a Carlos su comprensión y apoyo incondicional en cada uno de mis proyectos.

Granada, Abril de 2013

Smoothing with B-spline bases

1.1 Introduction

Functional data analysis (FDA) is a topic of active statistical research devoted mainly to the extension of multivariate analysis techniques to the case where data consist of a set of curves instead of vectors. A functional data set provides information about functions (curves, surfaces, etc) varying over a continuum. In most of applications, sample curves come from the observation of a stochastic process in continuous time.

The argument of the sample functions is often time, but may also be a different magnitude such as spatial location, wavelength or probability. In spectroscopy, for example, the NIR spectrum is a functional variable whose observations are measured as functions of wavelengths. The potential of functional data analysis methodologies for the chemometric analysis of spectroscopic data was shown in Saeys et al. (2008). A magistral compilation of models working with sample curves and interesting applications in different fields are collected in Ramsay and Silverman (2005) and Ramsay and Silverman (2002), respectively. A part of the literature has recently been concerned with functional data in a wide variety of statistical problems, and with developing procedures based on smoothing techniques.

Despite their continuous nature, sample curves are usually observed in a

finite set of sampling points that could be unequally spaced and different among the sample units. Because of this it is necessary to reconstruct the true functional form of each sample curve from a finite set of discrete observations. Many approximation techniques such as interpolation or projection in a finite-dimensional space generated by basis functions were applied from the beginning to solve this problem. More recently, nonparametric techniques were used for approximating functional data (Ferraty and Vieu, 2006).

In many applications the data are smooth functions observed with error. In this case least squares approximation can be used to estimate the basis coefficients of a basis expansion of the unobserved smooth sample functions. In this chapter three different approaches for solving this problem in terms of B-spline basis functions are compared in the FDA context: regression splines, smoothing splines and penalized splines.

B-splines are constructed from polynomial pieces joined at a set of knots. Once the knots are given, B-splines can be evaluated recursively for any degree of the polynomial by using a numerically stable algorithm (see De Boor, 2001). The choice of knots is an important problem when working with B-splines. If too many knots are selected you have an overfitting of the data. On the other hand too few knots provides an underfitting. This fact is specially significant in the case of non-penalized spline regression (regression splines). Some automatic numerical schemes for optimizing the number and the position of the knots were proposed to solve this problem (see for example Friedman and Silverman, 1989).

Smoothing splines were first proposed by O'Sullivan (1986) by introducing a penalty in the second derivative of the curve. This approach restricts the flexibility of the fitted curve and prevents the overfitting. This approximation was generalized later such that it could be applied in any context where regression on B-splines was useful (Eilers and Marx, 1996). This kind of penalized smoothers known as P-splines work with a relatively large number of equally spaced knots and a penalty based on differences between coefficients of adjacent B-splines.

The approximation of smooth functions with B-spline bases is used in the estimation of a wide variety of FDA methodologies such as functional linear regression models, functional generalized linear models and functional additive models, among others (Brumback and Rice, 1998; Marx and Eilers, 1999; Cardot et al., 2003; Crambes et al., 2009; Aguilera et al., 2010b).

This justifies the importance of a comparison among the main smoothing approaches in terms of B-splines and to draw conclusions that allow the researchers and practitioners to use the most powerful tool in each case.

After this introduction section, a revision of the different non-penalized and penalized spline smoothers with B-spline bases (regression splines, smoothing splines and P-splines) is presented. The most used methods for choosing the smoothing parameter in the roughness penalty approaches (cross validation and generalized cross validation) are also adapted to select only one smoothing parameter for fitting all the sample curves in the FDA context. The comparison of the approximation results provided by the studied approaches is developed on a simulation study. Finally, the performance of these spline smoothers is also studied in two applications with real data.

1.2 Basic tools for FDA

In this section, the notation and the basic tools related to functional variables are summarized.

Let (Ω, \mathcal{A}, P) a probabilistic space and $(H, \langle \cdot, \cdot \rangle_H)$ a separable Hilbert space. Then, a Hilbertian random variable on H is defined as a measurable function

$$\begin{aligned} X : \Omega &\longrightarrow H \\ \omega &\longrightarrow X(\omega), \end{aligned}$$

such that $X^{-1}(B) \in \mathcal{A}$, being B a Borel set of the Borel σ -algebra generated by the topological space H .

In this thesis, we are focus on functional variables whose observations are realizations of a continuous-time stochastic process $\{X(t) : t \in T\}$, with sample functions in the Hilbert space $L^2(T)$ of integrable square functions on T defined by

$$L^2(T) = \left\{ f : T \longrightarrow \mathbb{R} : \int_T f^2(t) dt < \infty \right\},$$

with the usual scalar product given by

$$\langle f, g \rangle = \int_T f(t) g(t) dt, \quad \forall f, g \in L^2(T). \quad (1.1)$$

Working with stochastic processes it is usual to assume that they are second order stochastic processes. A random process $\{X(t) : t \in T\}$ is a second

order stochastic process if $\forall t \in T$ the random variable $X(t) \in L^2(\Omega)$, where $L^2(\Omega)$ is the space of real random variables X on Ω with finite second order moment, so that

$$E[|X|^2] = \int_{\Omega} |X(\omega)|^2 dP(\omega) < \infty, \quad \forall X \in L^2(\Omega).$$

Let us consider on $L^2(\Omega)$ the natural scalar product defined by

$$\begin{aligned} L^2(\Omega) \times L^2(\Omega) &\longrightarrow \mathbb{R} \\ (X, Y) &\longrightarrow E[XY] = \int_{\Omega} X(\omega)Y(\omega)dP(\omega). \end{aligned}$$

Then, $L^2(\Omega)$ with the natural scalar product defined above has structure of Hilbert space.

Associated with a second order stochastic process, let us define the following functions essential for the development of the methodologies presented in this thesis:

- Mean function

$$\begin{aligned} \mu : T &\longrightarrow \mathbb{R} \\ t &\longrightarrow \mu(t) = E[X(t)] = \int_{\Omega} X(t, \omega) dP(\omega). \end{aligned}$$

- Covariance function

$$\begin{aligned} C : T \times T &\longrightarrow \mathbb{R} \\ (t, s) &\longrightarrow C(t, s), \end{aligned}$$

where

$$\begin{aligned} C(t, s) &= E[(X(t) - \mu(t))(X(s) - \mu(s))] \\ &= \int_{\Omega} [(X(t, \omega) - \mu(t))(X(s, \omega) - \mu(s))] dP(\omega). \end{aligned}$$

Most of the functional techniques impose certain restrictions regarding the continuity of the covariance function. Then, the continuity in quadratic mean of the stochastic process is required, involving the continuity of covariance function in $T \times T$.

A real stochastic process $\{X(t) : t \in T\}$ is continuous in quadratic mean if

$$\lim_{h \rightarrow 0} E[(X(t+h) - X(t))^2] = 0, \quad \forall t \in T.$$

On the other hand, it is known that if a process is continuous in quadratic mean, there is another process stochastically equivalent, whose sample paths are integrable square functions. Then, hereinafter it will be considered a functional random variable X whose observations are realization of a stochastic process $\{X(t) : t \in T\}$ verifying the following hypothesis:

H_1 : $\{X(t) : t \in T\}$ is the second order

H_2 : $\{X(t) : t \in T\}$ is continuous in quadratic mean

H_3 : The sample paths belong to $L^2(T)$.

Then, a continuous stochastic process may be seen as a random function defined on $L^2(T)$:

$$\begin{aligned} X : \Omega &\rightarrow L^2(T) \\ \omega &\rightarrow X(\omega) : T \rightarrow \mathbb{R} \\ & \quad t \rightarrow X(t, \omega). \end{aligned}$$

Associated with a stochastic process, under the hypothesis H_1 , H_2 and H_3 , the covariance operator is defined as

$$\begin{aligned} \mathcal{C} : L^2(T) &\longrightarrow L^2(T) \\ f &\longrightarrow \mathcal{C}(f)(t) = \int_T C(t, s) f(s) ds. \end{aligned}$$

Let us observe that the covariance function $C(t, s)$ is the kernel of the covariance operator \mathcal{C} . Then, as C is a continuous function in $T \times T$, \mathcal{C} is a bounded and continuous operator in the Hilbert space $L^2(T)$.

As in the multivariate case, in order to study a phenomena, a sample of observations is required. Hereinafter, let us consider a random sample of size n of a functional variable X denoted by $\{x_i(t) : t \in T, i = 1, \dots, n\}$. The sample paths can be considered as independent and equally distributed realizations of a continuous second order stochastic process $X = \{X(t) : t \in T\}$. Then, the sample mean function is given by

$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t) \quad \forall t \in T,$$

and the sample covariance function will be

$$\hat{C}(s, t) = \frac{1}{n-1} \sum_{i=1}^n (x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)), \quad \forall s, t \in T.$$

These functions are unbiased and consistent estimators that converge almost surely to the corresponding population moments (Deville, 1974).

1.3 Basis expansion of functional data

As indicated above, the first step in FDA is to reconstruct the functional form of the sample curves from their discrete observations. The most usual way to solve this problem consists of assuming an expansion of each sample curve in terms of a basis of functions and to fit the basis coefficients using smoothing or interpolation.

Let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of functions which is the sample information related to a functional variable X .

In practice, sample functions are observed in a finite set of time points $\{t_{i0}, t_{i1}, \dots, t_{im_i} \in T\} \forall i = 1, \dots, n$. Then, the sample information is given by the vectors $x_i = (x_{i0}, \dots, x_{im_i})'$, with x_{ik} being the value of the i -th sample path $x_i(t)$ observed at the time t_{ik} ($k = 0, \dots, m_i$). Because of this, the first step in FDA is to get the functional form of the sample curves.

In this section, the sample paths are assumed to belong to a finite-dimension space generated by a basis $\{\phi_1(t), \dots, \phi_p(t)\}$ so that they are expressed as

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad i = 1, \dots, n. \quad (1.2)$$

This equation can be expressed in matrix form as $x_i(t) = a_i' \phi(t)$, where $a_i = (a_{i1}, \dots, a_{ip})'$ and $\phi(t) = (\phi_1(t), \dots, \phi_p(t))'$.

There are different ways of obtaining the basis coefficients depending on the kind of observations we are working with. If the sample curves are observed without error

$$x_{ik} = x_i(t_{ik}) \quad k = 0, \dots, m_i, \quad i = 1, \dots, n,$$

some interpolation method, such as natural cubic spline interpolation, can be used (Aguilera et al., 1996). Quasi-natural cubic spline interpolation with B-splines functions was used to reconstruct sample curves of temperatures from daily observations and to predict the annual risk of drought in terms of them (Escabias et al., 2005). If the functional predictor is observed with error

$$x_{ik} = x_i(t_{ik}) + \varepsilon_{ik} \quad k = 0, \dots, m_i, \quad i = 1, \dots, n, \quad (1.3)$$

we can use a smooth approximation method as least squares after choosing an appropriate basis. An application of least squares smoothing with trigonometric and B-spline basis was developed for approximating the curves of stress of lupus patients from daily observations and determining the relationship between flares and stress level (Aguilera et al., 2008a).

With both methods, smoothing and interpolation, the functional form of sample paths is obtained by approximating the basis coefficients $\{a_{ij}\}$ from the observations of the sample curves at discrete points.

The goal is fitting a function x_i from each vector $x_i = (x_{i1}, x_{i2}, \dots, x_{im_i})'$ of discrete noisy observations by assuming model (1.3) and basis expansion (1.2) for each one of the n observed sample curves.

Choosing the ideal basis and its dimension p for approximating the functional form of a set of sample curves is very important and must be done according to the characteristics of the data. Useful basis systems are Fourier basis for periodic data, B-spline basis for non-periodic smooth data with continuous derivatives up to certain order, and wavelet basis for data with a strong local behavior whose derivatives are not required. In this thesis, smooth sample curves observed with error will be considered. Because of this, different types of least squares smoothing with spline functions are revised and compared. B-spline basis functions that have excellent numerical properties are considered to span the spline smoothers. The study of spline functions from an introductory level to a higher mathematical level can be followed in Green and Silverman (1994), De Boor (2001) and Wahba (1990), respectively. Recently, basis of splines were used to evaluate paper manufactured using *Eucalyptus globulus* by means of multivariate adaptive regression splines (García Nieto et al., 2012). Formulas and computational algorithms for optimal smoothing curves with B-splines basis for given set of discrete data, not necessarily equally spaced data, were studied in detail in Kano et al. (2005, 2011).

1.3.1 B-splines

A B-spline basis of degree q (order $q+1$) generates the space of the splines of the same degree, defined as curves consisting of piecewise polynomials of degree q that join up smoothly at a set of definition knots with continuity in their derivatives up to order $q-1$. In Figure 1.1 (top) an example of B-spline

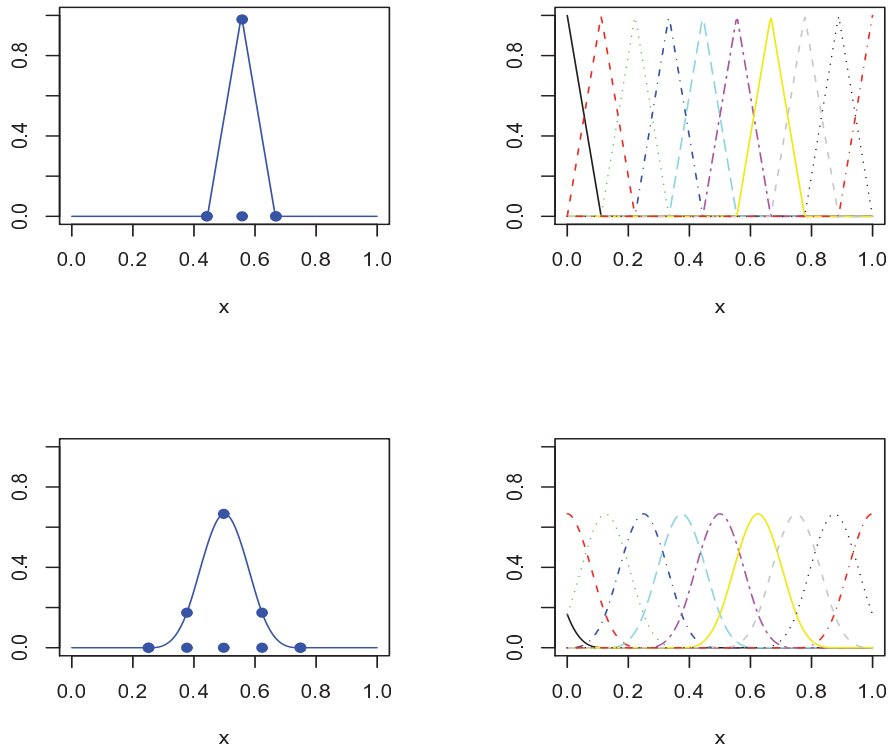


Figure 1.1: B-spline bases of order 2 (top) and 4 (bottom).

basis of order 2 is shown. A B-spline basis of order 4 is displayed at the bottom of the same Figure.

The dimension of the B-spline basis of degree q equals the order of the polynomials plus the number of interior breakpoints (see De Boor (2001) for a detailed explanation). The spline functions of degree q are smooth and well-behaved functions that provide design flexibility so that by increasing the degree q , we can progressively switch from the simplest piecewise constant ($q = 0$) and piecewise linear ($q = 1$) representations to the other extreme, which corresponds to a bandlimited model ($n \rightarrow \infty$).

Let $\tau_0 < \dots < \tau_s$ be a partition of knots of the observation interval T . Extending the partition as $\tau_{-3} < \tau_{-2} < \tau_{-1} < \tau_0 < \dots < \tau_s < \tau_{s+1} < \tau_{s+2} < \tau_{s+3}$, the B-spline basis of order $q+1$ is iteratively defined by (De Boor, 1977)

$$B_{j,1}(t) = \begin{cases} 1 & \tau_{j-2} \leq t < \tau_{j-1} \\ 0 & \text{in other case} \end{cases}, \quad j = -1, 0, 1, \dots, s+4$$

$$B_{j,q+1}(t) = \frac{t - \tau_{j-2}}{\tau_{j+q-2} - \tau_{j-2}} B_{j,q}(t) + \frac{\tau_{j+q-1} - t}{\tau_{j+q-1} - \tau_{j-1}} B_{j+1,q}(t)$$

$$q = 1, 2, \dots; \quad j = -1, 0, \dots, s - q + 4.$$

When $q = 3$ this basis functions are called cubic B-splines. They are used to fit regular sample curves with first and second continuous derivatives. From now on, the subscript corresponding to the order of B-spline basis functions will be omitted so that cubic B-splines will be denoted as

$$B_{j,4}(t) = B_j(t), \quad j = -1, 0, \dots, s+1.$$

Let us observe that the dimension of the B-spline basis of degree q equals the order of the polynomials plus the number of interior breakpoints. Then, the dimension of the cubic B-spline basis with knots $\tau_0 < \dots < \tau_s$ is the total number of knots plus two ($s+3$).

1.3.2 Other bases

As we said before, there are other useful bases. The most common ones are summarized hereinafter:

1. Fourier basis for periodic data

The orthonormal version of the Fourier basis is known as orthonormal basis of trigonometric functions in $L^2(T)$ and is given by

$$\begin{aligned} \phi_0(t) &= \frac{1}{T^{1/2}} \\ \phi_{2j-1}(t) &= \left(\frac{2}{T}\right)^{1/2} \sin\left(\frac{2\pi jt}{T}\right) \\ \phi_{2j}(t) &= \left(\frac{2}{T}\right)^{1/2} \cos\left(\frac{2\pi jt}{T}\right) \quad j = 1, \dots \end{aligned}$$

The basis representation with this kind of basis is known as Fourier development. This type of basis is accurate when working with periodic data. A successful application of FPCA with Fourier basis expansions was developed in Valderrama et al. (2010) where different functional principal component regression models were developed to forecast cypress pollen concentration from daily evolution of temperatures.

2. Wavelet bases

A basis of wavelets is obtained by considering dilations and translations of a suitable mother wavelet ϕ ,

$$\phi_{kj}(t) = 2^{k/2} \phi(2^k t - j),$$

with j and k being integers.

The wavelet expansion provides a decomposition of a function into orthogonal signal components at different resolution levels that it is called multiresolution analysis. The advantages of this wavelet representation derive from the ability of wavelets to represent locally non-smooth functions with only a relatively small number of coefficients. Because of this, wavelet analysis provide useful methods for analyzing data with intrinsically local properties, such as discontinuities and sharp spikes. A recent study on the estimation of multidimensional curves and their derivatives by using wavelets was developed by Pigoli and Sangalli (2012).

3. Polynomial bases

This functions are rarely used because despite its ease of calculation, they have large oscillations and lack of smoothness. A polynomial basis can be expressed as

$$\phi_j(t) = (t - \theta)^j, \quad j = 0, 1, 2, \dots$$

with θ being a shift parameter that is usually chosen to be in the center of the interval of approximation. These functions do not show an accurate local behavior unless the degree will be high. Moreover, polynomials also tend to fit well in the center and quite bad in the queues.

4. Constant basis

The sample path associated with point and counting processes are constant at random intervals defined by the instants at which new arrivals

occur. An appropriate basis for reconstructing the sample functions of such processes is the orthogonal basis of constant functions on the intervals of a fixed partition.

Given a partition of the observation interval T defined by the knots $0 = a_0 < a_1 < \dots < a_p$, an orthonormal basis of the subspace of constant functions over each of the intervals $(a_{j-1}, a_j]$ ($j = 1, \dots, p$) is defined as

$$\phi_j(t) = (a_j - a_{j-1})^{-1/2} I_t(t),$$

where $I_t(t)$ is the indicator function in the interval $(t_{j-1}, t_j]$ that takes the value 1 in this interval and zero outside of it.

1.4 Smoothing with B-spline bases

Different ways of approximating the basis coefficients in terms of B-spline bases are reviewed in this section. In Durban (2009) a comparative study of regression splines and smoothing splines was carried out. A complete guide about the use of splines with penalty (P-splines) in different models can be seen in Durban (2007).

1.4.1 Regression splines

Let us consider the basis expansion of the sample paths given by Equation (1.2), which can be expressed in matrix form as $x_i(t) = a_i' \phi(t)$, with $a_i = (a_{i1}, \dots, a_{ip})'$.

The simplest linear smoother approximates the coefficients a_i by minimizing the least squares criterion

$$MSE(a_i | x_i) = (x_i - \Phi_i a_i)' (x_i - \Phi_i a_i),$$

with $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$.

Thus, the estimate of a_i that minimizes this mean squared error is given by

$$\hat{a}_i = (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i.$$

Then, fitted values at the observation knots are given by the vectors

$$\hat{x}_i = \Phi_i \hat{a}_i = \Phi_i (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i,$$

and the fitted curves by

$$\hat{x}_i(t) = \hat{a}'_i \phi(t) \quad \forall i = 1, \dots, n.$$

When a basis of B-splines is considered, these fitted curves are usually called regression splines.

This approximation is appropriate when the errors ε_{ik} are independently distributed with zero mean and constant variance $\forall k = 0, \dots, m_i; i = 1, \dots, n$. In many applications with functional data the errors could be non-stationary and/or autocorrelated so that this assumption is not realistic. In these cases weighted least squares regression can be used (see a detailed study in Ramsay and Silverman, 2005).

The degree of smoothness of regression splines depends on the size of the B-spline basis which is a function of the number of knots and the degree of the spline. The choice of the number of knots is an important problem when working with regression splines because they do not control the degree of smoothness of the estimated curve. If too many knots are selected, you have an over-fitting of the data. On the other hand, too few knots provide an under-fitting. In Figure 1.2 it can be seen that the largest number of knots provides the worst fit to the underlying function because it does not filter out noise efficiently.

The selection of the number and location of knots in regression splines is through quite complicated and non attractive algorithms. See for example, Friedman and Silverman (1989); Lee (2000); Zhou and Shen (2001).

Localized smoothing methods such as kernel smoothing and local polynomial smoothing are an alternative class of weighted least squares smoothing with excellent computational properties but an important instability near the boundaries of the observational interval (Ferraty and Vieu, 2006). Continuous and discrete roughness penalty approaches are considered in this chapter as more flexible and powerful way of smoothing discrete data by a smooth function that solves the drawbacks of the ones mentioned before. In this case, the smoothness of the approximated curve is controlled by the smoothing parameter.

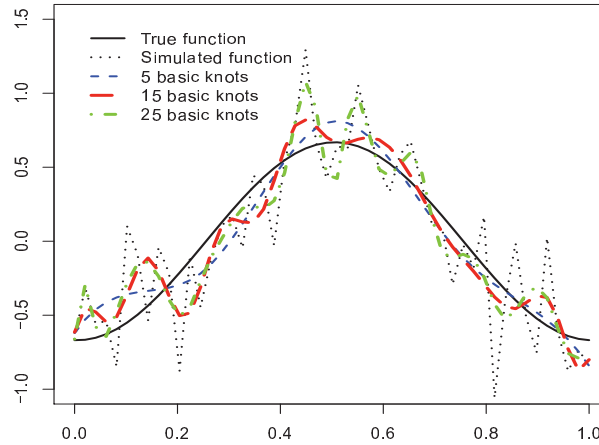


Figure 1.2: Regression spline approaches with different number of basis knots (5, 15 and 25).

1.4.2 Smoothing splines

Let us remember that the goal is to estimate for each sample curve the coefficients of its basis expansion from a set of discrete noisy observations that verify Equation (1.3). The curve fitted using roughness penalties provides a good fit to the data in terms of residual sum of squares and simultaneously controls the degree of smoothness.

The continuous penalty for smoothing splines measures the roughness of a function by means of the integrated squared second derivative and was first introduced by Reinsch (1967). If it is necessary, higher order of derivative can be used to control the degree of smoothness of the true curve. The computation of this continuous penalty in terms of B-splines basis functions was considered in O'Sullivan (1986) to propose optimal algorithms for solving the inverse problem.

In order to quantify the roughness of each curve $x_i(t)$, the integrated square of the d-order derivative is considered

$$\int [D^d x_i(s)]^2 ds = a_i' R_d a_i,$$

where R_d is the matrix defined by

$$R_d = \int D^d \phi(s) D^d \phi(s)' ds, \quad (1.4)$$

with $D^d \phi(s) = (D^d \phi_1(s), \dots, D^d \phi_p(s))'$.

Then, the basis coefficients of the smoother are obtained by minimizing the continuous penalized least squares error given by

$$CPMSE_d(a_i|x_i) = (x_i - \Phi_i a_i)'(x_i - \Phi_i a_i) + \lambda a_i' R_d a_i. \quad (1.5)$$

In practice, the most common penalty order is $d = 2$. In this case it is well known (De Boor, 2001) that if the only assumption about the function is that the integral of its squared second derivative is finite, then the function that minimizes the penalized error given in Equation (1.5) is a cubic spline with knots at the data points $(t_{ik} : k = 0, \dots, m_i)$. This explains that the most common computational approach for spline smoothing is to minimize penalized criterion given in Equation (1.5) with respect to the coefficients of a basis expansion in terms of cubic B-splines functions with knots at the sampling points. In this case, the fitted function is called cubic spline smoother.

When a very large number of sampling points is involved a lower number of appropriate knots can be sufficient to smooth the sample paths and capture their main features. In general, a smoothing spline is obtained assuming an expansion in terms of B-splines and minimizing (1.5). Then, the vector of estimated basis coefficients is given by

$$\hat{a}_i = (\Phi_i' \Phi_i + \lambda R_d)^{-1} \Phi_i' x_i.$$

An interesting application of cubic smoothing splines for the implementation of the functional mixed effects models can be seen in Guo (2004).

1.4.3 P-splines

The roughness penalties considered for smoothing splines are defined in terms of integrated squared derivatives. The computational problem of this approach lies in the calculation of the matrix R_d whose elements are the integrals of products of d-order derivatives between B-spline basis functions. A simpler discrete penalty approach is based on defining the roughness of a function by summing squared d-order difference values. This kind of penalty

depends on the considered basis and only works if the sampling points are equally spaced. A penalty based on d -order differences between coefficients of adjacent B-splines is used in Eilers and Marx (1996). This type of smoothers are called penalized splines and they can be also computed in terms of truncated power functions. A recent study has shown that penalized spline regression in terms of B-splines with equally spaced knots and difference penalties outperforms the penalized spline approach based on truncated power functions with knots based on quantiles of the independent variable and a ridge penalty (Eilers and Marx, 2010).

The basis coefficients of a penalized spline smoother in terms of B-spline basis functions can be computed by minimizing the discrete penalized least squares error as follows

$$DPMSE_d(a_i|x_i) = (x_i - \Phi_i a_i)' (x_i - \Phi_i a_i) + \lambda a_i' P_d a_i, \quad (1.6)$$

where $P_d = (\Delta^d)' \Delta^d$ with Δ^d being the matrix of d -order differences given by the $(p-d) \times p$ matrix

$$\Delta^d = \begin{pmatrix} \binom{d}{0} & \binom{d}{1} & \binom{d}{2} & \cdots & \binom{d}{d} & 0 & 0 & \cdots \\ 0 & \binom{d}{0} & \binom{d}{1} & \cdots & \binom{d}{d-1} & \binom{d}{d} & 0 & \cdots \\ 0 & 0 & \binom{d}{0} & \cdots & \binom{d}{d-2} & \binom{d}{d-1} & \binom{d}{d} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (1.7)$$

Let us observe that the vector of d -order differences of the vector $a_i = (a_{i1}, \dots, a_{ip})'$ is given by $\Delta^d a_i$ and its components are the one-order differences of the vector of differences of order $d-1$ given by

$$\sum_{j=0}^d \binom{d}{j} a_{i, k+j} \quad k = 1, \dots, p-d.$$

The most common penalty matrix is $P_2 = (\Delta^2)' \Delta^2$, with Δ^2 the $(p-2) \times p$

matrix of 2-order differences given by

$$\Delta^2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots \\ 0 & 1 & -2 & 1 & 0 & \dots \\ 0 & 0 & 1 & -2 & 1 & \dots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \dots \end{pmatrix}.$$

These smoothers are called penalized splines (P-splines) and their B-spline basis coefficients are estimated by

$$\hat{\alpha}_i = (\Phi_i' \Phi_i + \lambda P_d)^{-1} \Phi_i' x_i. \quad (1.8)$$

The difference penalty is a good discrete approximation to the integrated square of the d -th derivative so that generalizations to penalties on higher derivatives can be easily implemented. On the other hand, P-splines are computed by penalizing the basis coefficients of the curves that reduces the dimensionality of the problem. We can say that P-splines combine the best of the regression splines and smoothing splines because they have less numerical complexity than smoothing splines and the selection of knots is not so determinant as in regression splines.

The application of P-splines to different models with smooth components (semi-parametric models, models with serially correlated errors, and models with heteroscedastic errors) and a nonparametric strategy for the choice of the P-spline parameters has been performed by Currie and Durban (2002), where mixed model (REML) methods were applied for smoothing parameter selection. Taking into account that the degree of smoothing is controlled by the smoothing parameter, the number and location of knots is not crucial for fitting a P-spline. Generally, the knots of a P-spline are equally spaced and the number of knots must be sufficiently large to fit the data and not so large that computation time is unnecessarily big. Two algorithms for automatic selection of the number of knots by using generalized cross validation were considered in Ruppert (2002).

1.5 Choosing the smoothing parameter

The role of the smoothing parameter in penalized smoothing is to control the smoothness of the fitted curve. As λ becomes larger the fitted function is smoother so that when $\lambda \rightarrow \infty$ the standard linear regression to the observed

data is implemented. On the other hand, when λ becomes smaller the fitted curve is more and more variable so that when $\lambda \rightarrow 0$ we have an interpolant to the data.

In order to compute the optimal value of the smoothing parameter λ , two selection criteria are considered and compared in this chapter: leave-one-out cross validation (CV) and generalized cross validation (GCV).

In order to select the same smoothing parameter for all the n sample paths we propose to minimize the mean of the cross validation errors over all sample curves.

The CV (leave-one-out) method consist of selecting, for each curve, the smoothed parameter λ that minimizes the next expression

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n CV_i(\lambda),$$

where

$$CV_i(\lambda) = \sqrt{\sum_{k=0}^{m_i} (x_{ik} - \hat{x}_{ik}^{-k})^2 / (m_i + 1)},$$

with \hat{x}_{ik}^{-k} being the values of the i -th sample path estimated at time t_{ik} avoiding the k -th time point in the iterative estimation process. The CV approach has two main problems, is very expensive from a computational point of view and can lead to under-smoothing the data.

The GCV method is computationally simpler and very used in the literature about smoothing splines (Craven and Wahba, 1978). We consider two versions of GCV error, one for the smoothing splines and other for P-splines. The GCV method consist of selecting λ so that minimize

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n GCV_i(\lambda),$$

where

$$GCV_i(\lambda) = \frac{(m_i + 1)^{-1} MSE_i}{[(m_i + 1)^{-1} tr(I - H_i)]^2}.$$

Equivalently, an easier way to interpret it would be

$$GCV_i(\lambda) = \left(\frac{(m_i + 1)}{(m_i + 1) - df(\lambda)} \right) \left(\frac{MSE_i}{(m_i + 1) - df(\lambda)} \right),$$

where

$$MSE_i = \frac{1}{n} \sum_{k=0}^{m_i} (x_{ik} - \hat{x}_{ik})^2,$$

and $df(\lambda) = tr(H_i)$, with $H_i = \Phi_i (\Phi_i' \Phi_i + \lambda R_d)^{-1} \Phi_i'$ in the case of the smoothing splines. If we work with P-splines $H_i = \Phi_i (\Phi_i' \Phi_i + \lambda P_d)^{-1} \Phi_i'$ (considering P_d instead of R_d).

1.6 Simulation study

The ability of P-splines, smoothing splines and regression splines to approximate smooth curves observed with noise is tested on simulated data. The simulated data set consists of 100 sample paths of a second order stochastic process with zero mean given by

$$X(t) = R \cos(2\pi t + \theta),$$

where R and θ are i.r.v with distributions Raileigh(σ), with $\sigma = 0.3$, and Uniform $[0, 2\pi]$, respectively. Noisy observations of the sample paths were simulated at $m = 51$ equally spaced knots in the interval $T = [0, 1]$. That is,

$$x_{ik} = X(t_{ik}) + \epsilon_{ik} \quad (t_{ik} = k \times 0.02; k = 0, 1, \dots, 50; i = 1, \dots, 100),$$

where the errors ϵ_{ik} were simulated from independent normal distributions $N(0, \sigma^2)$ with $\sigma^2 = 0.07$ fixed to control the determination coefficient R^2 near 0.7.

The first step in this work was to select the smoothing parameter λ . In order to get the best smoothing parameter, we have compared the two different methods of selecting λ seen in Section (1.5). Figure 1.3 shows the box plot related to the mean squared error (MSE) of the approximated curves provided by the smoothing splines and P-splines with λ selected by CV and GCV. We can see that with the smoothing spline approach the CV method minimizes the MSE regardless the number of basis knots.

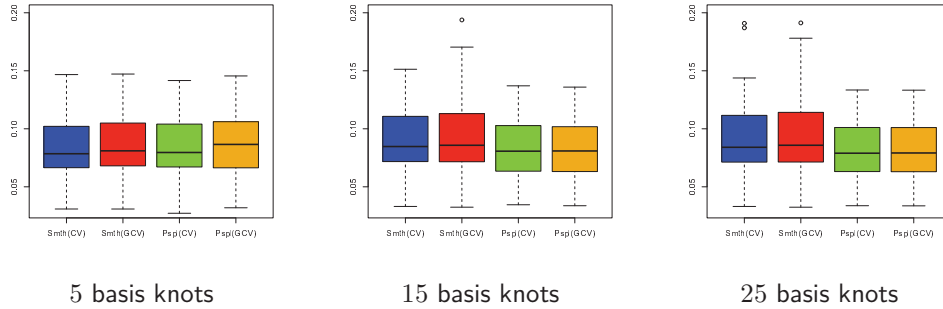


Figure 1.3: Simulation study. Box plot related to the MSE of the approximation of curves by smoothing splines, with λ selected by CV (blue) and GCV (red), and P-splines, with λ selected by CV (green) and GCV (orange), by using different number of basis knots (5, 15 and 25).

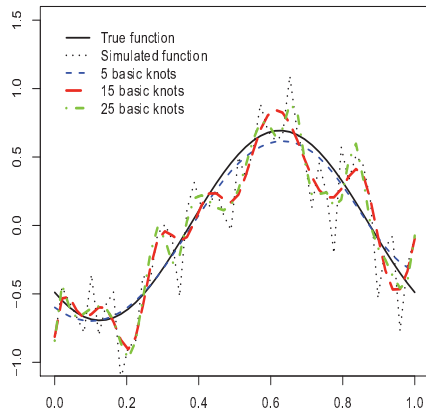
With the P-spline approach CV and GCV selection criteria provide similar approximation errors. In order to compare the three smoothing approaches with cubic B-spline bases studied in this work, the smoothing parameter was selected by CV method.

In Figure 1.4 and 1.5, the three different cubic spline approximations of a sample path to the simulated discrete data with different number of basis knots (5, 15 and 25) are displayed. It can be observed that the three smoothers are good approximations to the true function for the case of a 4-order B-spline basis with five knots. When the number of basis knots increases, regression splines and smoothing splines lose control of smoothness.

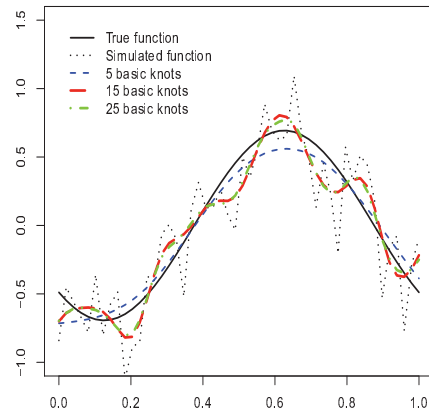
However, P-splines maintain a good fit for any number of knots. There are not too differences between smoothing splines and P-splines, being P-splines computationally easier to compute and its adjustment to the original function is not affected by the number of knots.

In order to obtain general conclusions, the mean curve and the MSE distribution provided by the three approximation approaches (regression splines, smoothing splines and P-splines) for the 100 simulated sample paths, have been represented in Figure 1.6 by considering different number of basis knots (5, 15 and 25).

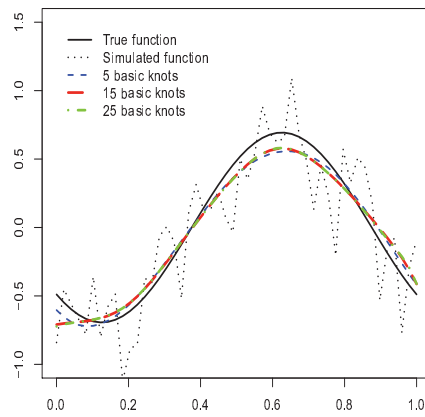
It can be seen that the P-spline approach provides the best fit to the true mean function and the smallest MSE in all considered cases. On the contrary, regression splines give the worst fit because they do not control the degree of smoothness.



Regression splines



Smoothing splines



P-splines

Figure 1.4: Simulation study. Regression splines, smoothing splines and P-splines approaches with different number of basis knots (5, 15 and 25) for one of the simulated sample curves.

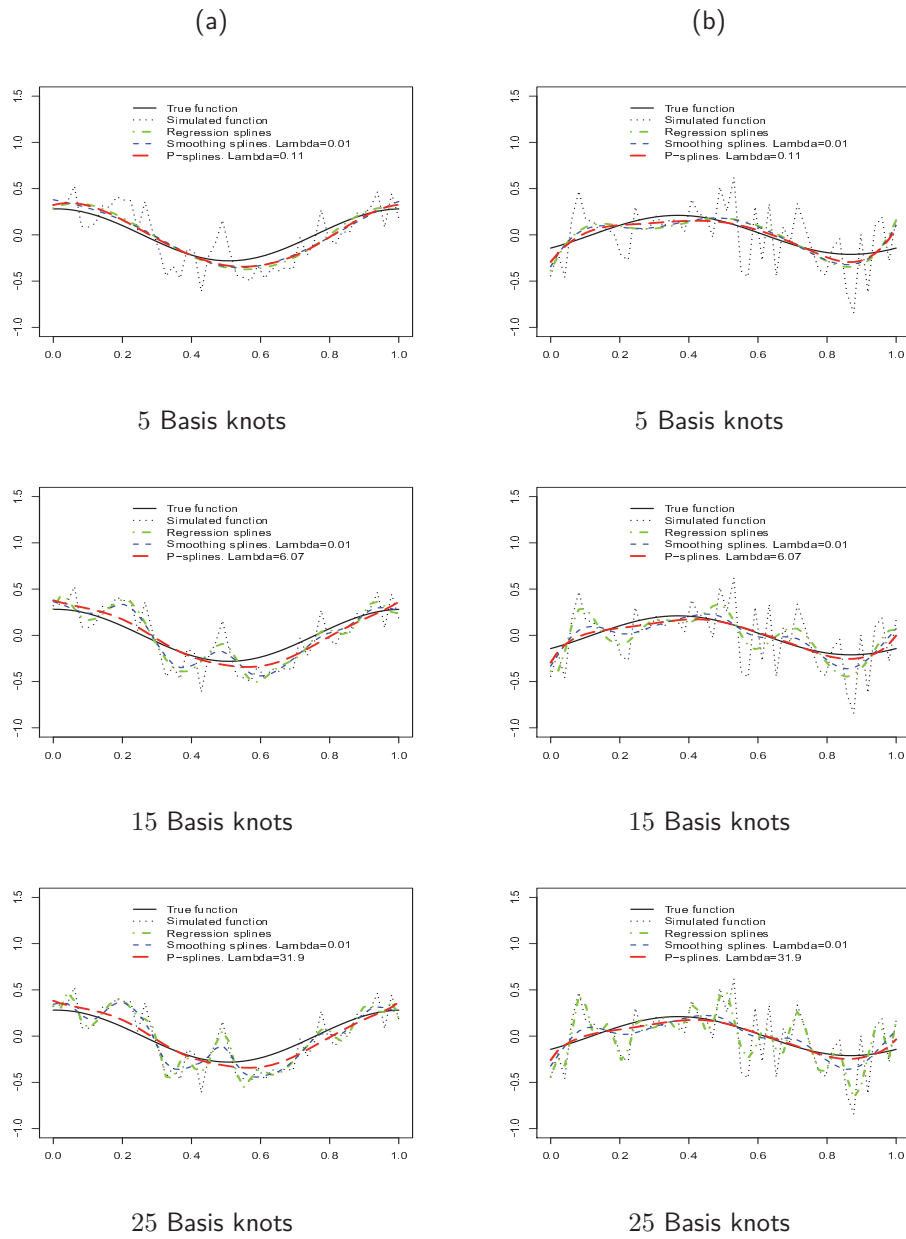
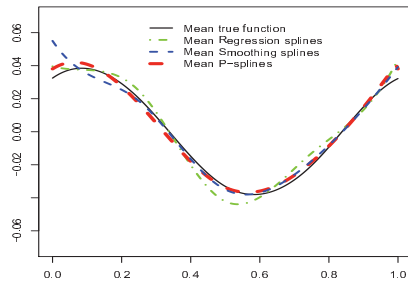
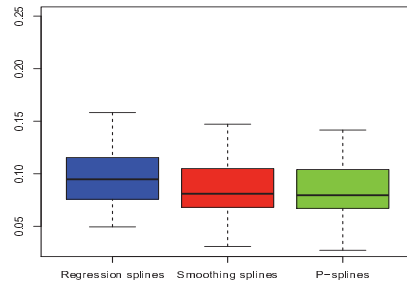


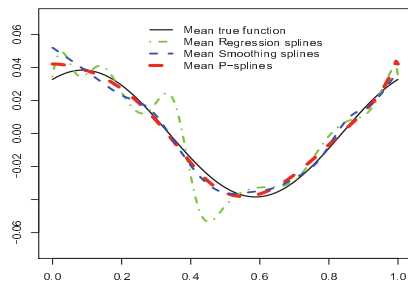
Figure 1.5: Simulation study. Regression splines (green, dashed and dotted line), smoothing splines (blue and dashed line) and P-splines (red and long dashed line) approaches with 5, 15 and 25 basis knots, for two different sample paths (a) (left) and (b) (right).



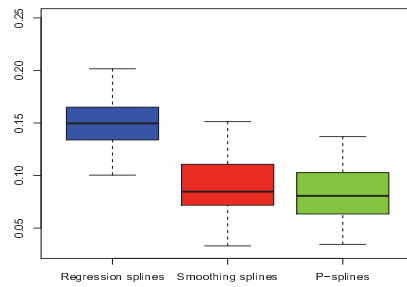
5 Basis knots



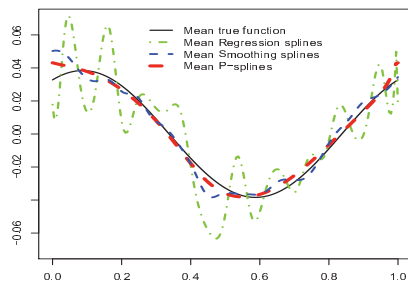
5 Basis knots



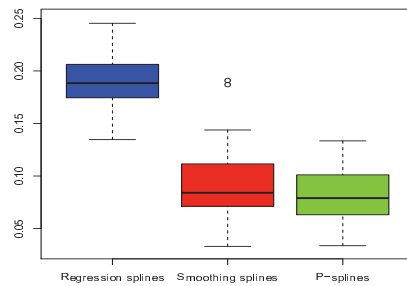
15 Basis knots



15 Basis knots



25 Basis knots



25 Basis knots

Figure 1.6: Simulation study. Mean function (left) and MSE (right) for 100 fitted curves through regression splines (blue), smoothing splines (red) and P-splines (green) approaches using 5, 15 and 25 basis knots. λ has been chosen by CV.

1.7 Real data applications

Once the P-splines have been chosen as the best smoothers to approximate noisy sample paths from discrete observations, their behavior have been tested using two real functional data sets. Firstly, we approximate the pinch force data set by using P-splines and comparing the results with the other two methodologies summarized in this chapter. In the second application, P-splines approach is applied to smooth the spectrometric curves related to Flemish hog manures.

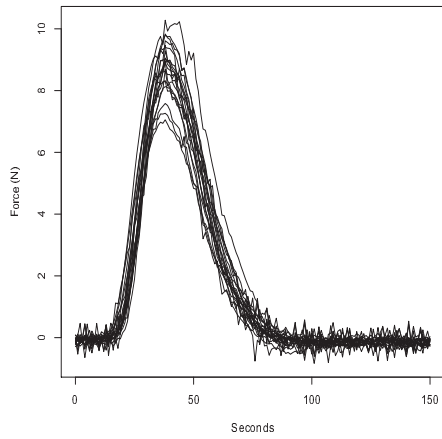
1.7.1 Pinch data

Pinch data, analyzed in Ramsay and Wang (1995), were collected at the Medical Research Council Applied Psychology Unit, Cambridge, and consist of records of the force exerted by pinching a force meter (width 6 cm) with the tips of the thumb and forefinger on opposite sides.

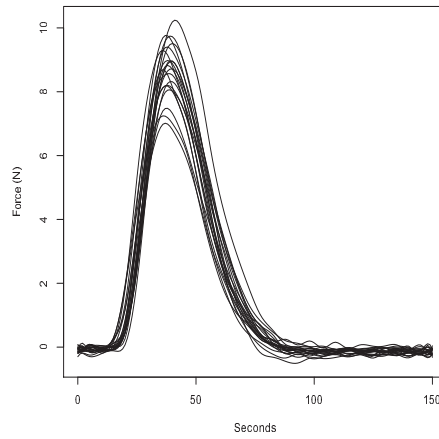
The exerted force must be adapted to the characteristics of the gripped object (such as texture, weight, surface, acceleration, between others). Sometimes, the system is slow to the response speed required by the exterior world and in this case, it is the brain who must exert the required force. So, the importance of studying this system is to make possible a better understanding of how the brain can control high performance motor systems.

The data set used in this chapter consists of a sample of 20 records of the force exerted by the human thumb and forefinger during a brief squeeze. The force was sampled at 151 times (seconds). We have considered a cubic B-spline basis with 30 equally spaced knots to approximate the true sample paths. The smoothing parameter λ has been chosen by CV method.

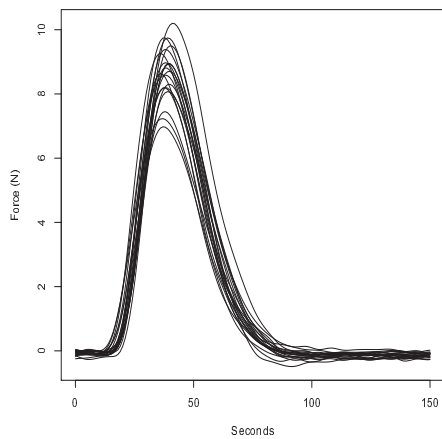
In Figure 1.7 (a) the necessity of smoothing the observed data is clear. The different spline approaches with B-spline basis studied in this work have been applied and display in Figure 1.7 (b), (c) and (d). Let us observe that regression splines can not completely avoid the noise at the extremes. Between the two kind of penalty applied (smoothing splines (c) and P-splines (d)), is the P-spline approach who provides the best smoothing of the sample paths. Two original sample paths and their P-spline approaches are shown in Figure 1.8.



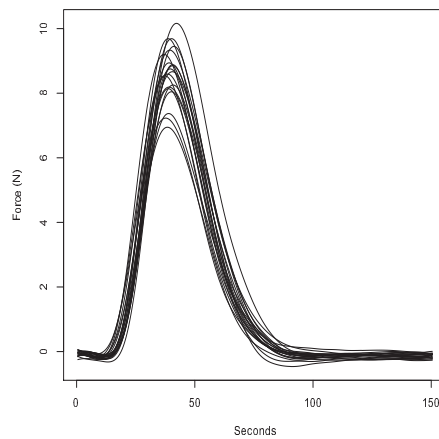
(a) Original sample paths



(b) Regression splines



(c) Smoothing splines



(d) P-splines

Figure 1.7: Application (pinch data). Original pinch data set (a) and its fit by regression splines (b), smoothing splines (c) and P-splines (d) using B-splines basis defined at 30 knots. The different values of λ have been chosen by CV.

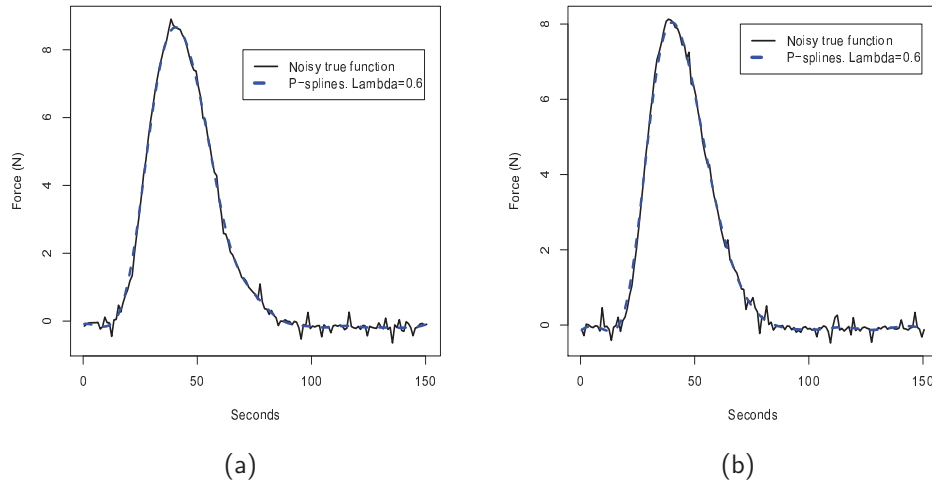
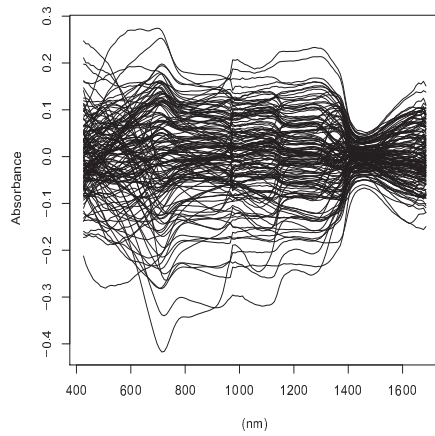


Figure 1.8: Application (pinch data). Fitting two true functions observed with noise, (a) and (b) (black and solid line) by P-splines (blue and dashed line) using 30 basis knots and $\lambda = 0.6$ (chosen by CV).

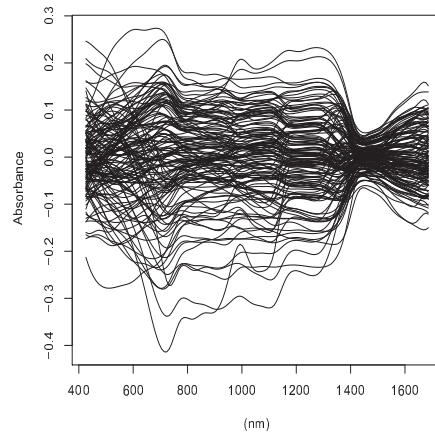
1.7.2 Manure data

Manure data set was analyzed in Saeys et al. (2004) and consists of 138 sample paths about Flemish hog manures collected in the spring of 2003 at almost as many different farms in Flanders by the Soil Service of Belgium. All samples were scanned in reflectance mode on a diode array Vis/NIR spectrophotometer. After that, data were converted into absorbance units ranging from 426 to 1686 nm.

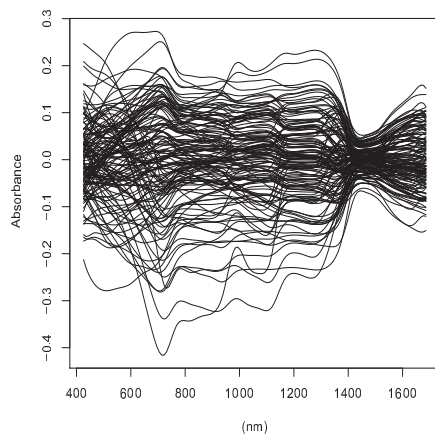
In order to compute the three different types of spline smoothers to the observed data, a cubic B-splines basis defined at 30 knots has been considered. The smoothing parameters λ have been chosen by CV method. The original sample paths are represented in Figure 1.9 (a). The smoothing splines (c) and regression splines (b) are quite similar. However, P-splines leads to the smoothest approximation of the sample paths. Finally, two original sample paths and their P-spline approximations are shown in Figure 1.10.



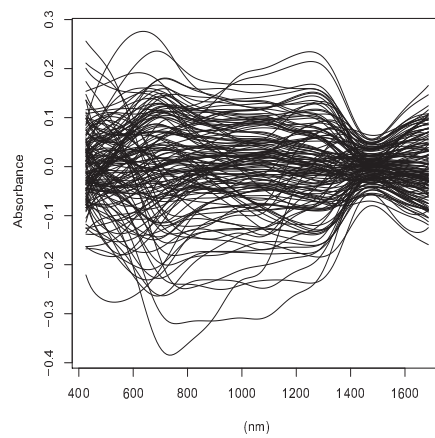
(a) Original sample paths



(b) Regression splines



(c) Smoothing splines



(d) P-splines

Figure 1.9: Application (manure data). Original manure data set (a) and its fit by regression splines (b), smoothing splines (c) and P-splines (d) using B-splines basis defined at 30 knots. The different values of λ have been chosen by CV.

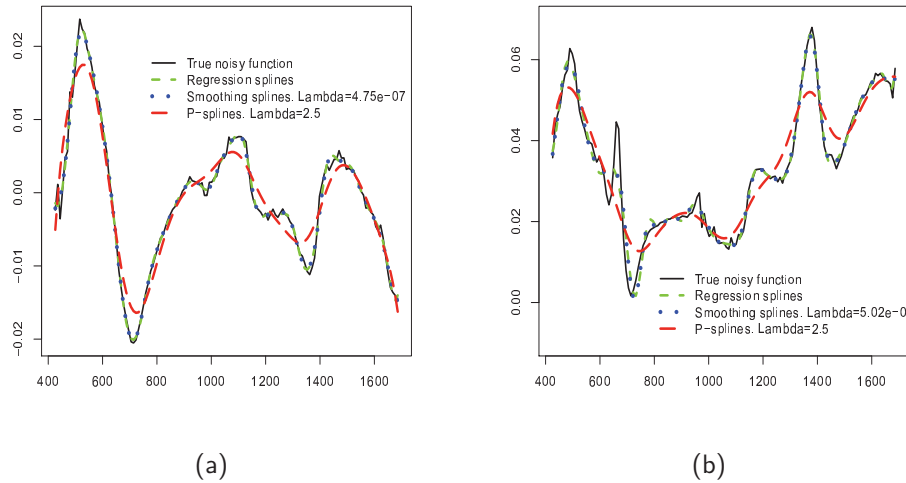


Figure 1.10: Application (manure data). Fitting two true functions observed with noise, (a) and (b) (black and solid line) by using P-splines (red and large dashed line), smoothing splines (blue and dotted line) and regression splines (green and dashed line), with 30 basis knots. λ chosen by CV.

1.8 Conclusions

Non-penalized and penalized least squares smoothing in terms of B-spline bases have been compared in this chapter to approximate a set of unobserved smooth curves from discrete noisy observations. A simulation study and two applications with real functional data have been developed to study and compare the performance of the three considered smoothers (regression splines, smoothing splines and P-splines) in the FDA context.

In base to these results we can conclude that regression splines and smoothing splines lose control of the smoothness when the number of knots increases. Both penalized approaches get to improve the fit providing mean squared errors with respect to the original smooth sample curves much smaller than the ones given by the non-penalized approach. On the other hand, P-splines provide the lowest approximation errors, having less numerical complexity that makes easier its computational implementation. Moreover, P-splines are quite insensitive to the choice of knots being sufficient to choose a relatively large number of equally spaced basis knots.

Penalized PCA approaches for B-spline expansions of smooth functional data

2.1 Introduction

When analyzing a functional data set it is usual to have a large number of regularly spaced observations for each sample curve. Because of this a reduction dimension technique is necessary for explaining the main features of a set of sample curves in terms of a small set of uncorrelated variables. This problem was solved by generalizing principal component analysis to the case of a continuous-time stochastic process (Deville, 1974). Asymptotic properties of the estimators of FPCA were deeply studied in the general context of functional variables (Dauxois et al., 1982). Nonparametric methods were developed to perform FPCA for the case of a small number of irregularly spaced observations of each sample curve (James et al., 2000; Yao et al., 2005a). As in the multivariate case, the interpretation of the principal component scores and loadings is a useful tool for discovering the relationships among the variables associated to a functional data set. To avoid misinterpretation of PCA, a new type of plots, named Structural and Variance Information plots, were recently introduced by Camacho et al. (2010).

FPCA is a flexible tool in functional data analysis that is successfully used to solve important problems as the estimation of the functional parameter in different functional regression models (Cardot et al., 1999; Aguilera et al., 1999; Cuevas et al., 2002; Escabias et al., 2004; Müller and Stadtmüller, 2005; Yao et al., 2005b; Cai and Hall, 2006). An alternative methodology for solving this estimation problem in the functional linear model is the functional version of partial least squares (PLS) regression. A estimation procedure for functional PLS based on basis expansions of sample curves was introduced by Aguilera et al. (2010b). A Bayesian approach to FPCA based on a generative model for noisy and sparse observations of curves was developed in Van der Linde (2008). Robust estimators for the functional principal components are considered by using basis expansion (Locantore et al., 1999) and by adapting the projection pursuit approach to the functional data context (Bali et al., 2011).

One usual form of estimating FPCA from discrete observations of the sample curves is based on basis expansion approximation. This way, FPCA of a set of curves is reduced to multivariate PCA of a transformation of the matrix of basis coefficients (Ocaña et al., 2007). B-spline bases are appropriate to approximate smooth curves. Cubic spline interpolation with B-spline basis can be considered for approximating smooth sample curves observed without error (Aguilera et al., 1996). Monotone piecewise cubic interpolation of the sample paths was proposed to approximate the mean of a doubly stochastic Poisson process in Bouzas et al. (2006). On the other hand, least squares approximation with B-spline basis is appropriate for reconstructing the true functional form of noisy smooth curves. This type of approximation was performed to forecast lupus flares from time evolution of stress level (Aguilera et al., 2008a). As it was shown in Chapter 1, the problem is that regression splines do not control the degree of smoothness and, consequently, the principal components are difficult to interpret because the estimated weight functions have a lot of variability and lack of smoothness. This problem must be solved by introducing some kind of smoothing in the estimation of principal component curves. A kernel approach based on regularizing the trajectories is considered in Boente and Fraiman (2000) to provide smooth estimators in FPCA.

There are different ways of introducing smoothing in the estimation of FPCA. On the one hand, the data can be smoothed first and then a non-penalized FPCA is carried out. A spline smoothing that penalizes the in-

egrated squared second derivative of each sample path was considered by Besse and Ramsay (1986) and Besse et al. (1997). This approach was applied for smoothing and reconstructing a magnetic resonance imaging (fMRI) functional data in Viviani et al. (2005). On the other hand, the smoothing can be introduced within the FPCA algorithm. Two different approaches for smoothing functional principal components analysis were proposed by Rice and Silverman (1991) and Silverman (1996). Both approaches use a continuous penalty that measures the roughness of the principal component curves by their integrated squared d -order derivative but they differ in the way they incorporate the penalty. The Rice-Silverman approach introduces the roughness penalty in the definition of the sample variance of the principal component weight functions. The Silverman approach is known as regularized FPCA (RFPCA) and introduces the penalty in the orthonormality constraint between principal components. This FPCA approach was extended to the case of multivariate functional data sets by using Gaussian basis functions instead of B-splines (Kayano and Konishi, 2009). An application of regularized FPCA with B-splines basis in actuarial science was performed to estimate the risk of occurrence of a claim in terms of the driver's age and others significative variables (Segovia-Gonzalez et al., 2009). A third way of penalizing FPCA is based on smoothing not the data or the components, but the covariance operator, whose eigenfunctions are the principal component functions (Yao et al., 2005a). Penalized rank one approximation was recently proposed as an alternative approach to the estimation of FPCA (Huang et al., 2008). On the other hand, equivalences between functional PCA with a proposed inner product and certain PCA with a given well-suited inner product were studied by Ocaña et al. (1999) in the theoretical framework given by Hilbert valued random variables, in which multivariate and functional PCAs appear jointly as particular cases.

Penalized spline regression (Eilers and Marx, 1996) is an increasingly popular smoothing approach that was used to estimate the functional sample mean and to develop an iterative P-spline algorithm for estimating FPCA in Yao and Lee (2006). The P-spline penalty measures the roughness of a function by summing squared d -order differences between adjacent basic coefficients. In this chapter, two different versions of smoothed FPCA based on penalized splines (P-splines) with B-splines basis are introduced and compared. The first approach carries out a non-penalized FPCA on the P-spline smoothing of the sample curves. The second approximates the sample curves by

non-penalized least squares and then incorporate the P-spline penalty in the orthonormality constraint within the FPCA algorithm. The accuracy of the estimates provided by both P-spline smoothed approaches is tested with simulated and real data, and the results compared with non-penalized FPCA and regularized FPCA. In order to get an optimum estimation of the smoothing parameter, leave-one-out cross validation is adapted to this context.

2.2 Functional principal component analysis

Let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of functions that are the sample information related to a functional variable X . It will be supposed that they are observations of a second order stochastic process $X = \{X(t) : t \in T\}$, continuous in quadratic mean whose sample functions belong to the Hilbert space $L^2(T)$ of square integrable functions with the usual scalar product defined in Equation (1.1).

Multivariate PCA was extended to the functional case to reduce the infinite dimension of a functional predictor and to explain its dependence structure by a reduced set of uncorrelated variables (Deville, 1974). In order to compute the functional principal components, let us assume without loss of generality that the observed curves are centered so that the sample mean $n^{-1} \sum_{i=1}^n x_i(t)$ is zero.

The principal components are obtained as uncorrelated generalized linear combinations with maximum variance. In general, the j -th principal component scores are given by

$$\xi_{ij} = \int_T x_i(t) f_j(t) dt, \quad i = 1, \dots, n. \quad (2.1)$$

where the weight function or loading f_j is obtained by maximizing the variance

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

It can be shown that the weight functions are the eigenfunctions of the sample covariance operator \hat{C} (defined in Chapter 1). That is, the solutions to the equation

$$\hat{C}(f_j)(t) = \int \hat{C}(t, s) f_j(s) ds = \lambda_j f_j(t), \quad (2.2)$$

where $\hat{C}(t, s)$ is the sample covariance function and $\lambda_j = \text{var}[\xi_j]$. We are considering that the sample curves comes from a second order stochastic process which is centered. Then, the Karhunen-Loève decomposition is considered, so that, the sample curves are expressed in terms of the functional principal components as follows:

$$x_i(t) = \sum_{j=1}^{n-1} \xi_{ij} f_j(t).$$

This principal component decomposition can be truncated providing the best linear approximation of the sample curves in the least squares sense

$$x_i^q(t) = \sum_{j=1}^q \xi_{ij} f_j(t),$$

whose explained variance is given by $\sum_{i=1}^q \lambda_i$.

The problem inherent to many applications is that interpreting the components is not always straightforward. This problem is usually solved by a rotation of the principal component curves that simplifies the factor structure and therefore makes the interpretation easier. There are two main types of rotation: orthogonal when the resulting factors are also orthogonal to each other, and oblique when the new factors are not required to be orthogonal to each other. Oblique rotations relax the orthogonality constraint in order to simplify the interpretation. They are used more rarely than their orthogonal counterparts but, recently, new techniques are developed based on oblique rotations. An example is independent component analysis that was originally created in the domain of signal processing and neural networks, and derives, directly from the data, an oblique solution that maximizes statistical independence (Hyvärinen et al., 2001).

The most popular orthogonal rotation method is indubitably Varimax. This rotation criterion has been applied to functional PCA in two different ways, one based on Varimax rotation of the matrix of basis coefficients of the principal component curves, and the other based on Varimax rotation of the matrix of values of the principal component curves in a grid of equally spaced time points (see Ramsay and Silverman (2005) for a detailed development of this functional Varimax rotation). The extension of oblique rotations to functional PCA is out of the scope of this work but could be developed by following the same idea that in the Varimax case.

2.2.1 Basis expansion estimation

In order to compute the principal component weights it is necessary to solve the second order integral Equation (2.2). It is a difficult problem that is further complicated in practice because the sample curves are usually observed at a finite set of sampling points that can be different for the sample individuals. This means that in real data applications the sample information is given by the vectors $\{x_i = (x_{i0}, \dots, x_{im_i})', i = 1, \dots, n\}$, with x_{ik} being the observed value for the sample path $x_i(t)$ at the sampling point $t_{ik}, k = 0, 1, \dots, m_i$.

The functional form of sample paths must be reconstructed from the discrete observations by using several different methods that must be chosen depending on how the functional data was observed and the main characteristics of the sample curves.

One usual solution is to assume that sample paths belong to a finite-dimension space spanned by a basis $\{\phi_1(t), \dots, \phi_p(t)\}$, so that they are expressed as in Equation (1.2).

The main objective of this chapter is to solve the problem of estimating the functional principal components from a sample of smooth curves observed with error so that the sample data are given by

$$x_{ik} = x_i(t_{ik}) + \epsilon_{ik} \quad k = 0, 1, \dots, m_i, \quad i = 1, \dots, n.$$

This means that least squares approximation with B-splines basis (De Boor, 2001) are an appropriate choice to approximate the basis coefficients.

Let us suppose that the sample paths are expressed in terms of basis functions, so that $x = A\phi$, where $A = (a_{ij})$ is the coefficient matrix, $\phi = (\phi_1, \dots, \phi_p)'$ and $x = (x_1, \dots, x_n)'$. Then, the principal component weight function f_j admits the basis expansion

$$f_j(t) = \sum_{k=1}^p b_{jk} \phi_k(t) = \phi(t)' b_j,$$

with $b_j = (b_{j1}, \dots, b_{jp})'$. In this case,

$$\begin{aligned} \text{var}[\xi] &= \text{var} \left[\int_T x_i(t) f(t) dt \right] \\ &= n^{-1} \sum_{i=1}^n \left[\int_T \left(\sum_{j=1}^p a_{ij} \phi_j(t) \right) \left(\sum_{k=1}^p b_k \phi_k(t) \right) \right] \\ &= n^{-1} \sum_{i=1}^n b' \Psi a'_i a_i \Psi b = b' \Psi V \Psi b, \end{aligned}$$

where $V = n^{-1} A' A$, with $A = (a_{ij})_{n \times p}$ and $\Psi = (\Psi_{ij})_{p \times p} = \int_T \phi_i(t) \phi_j(t) dt$. Therefore, computing the j -th principal weight function is reduced to solve the maximization problem

$$\begin{cases} \max_b b' \Psi V \Psi b \\ \text{s.t. } b' \Psi b = 1 \text{ and } b_\ell \Psi b = 0, \quad \ell = 1, \dots, j-1. \end{cases}$$

This means that FPCA is equivalent to the multivariate PCA of matrix $A \Psi^{\frac{1}{2}}$, with $\Psi^{\frac{1}{2}}$ being the squared root of the matrix of inner products between basis functions (Ocaña et al., 2007).

Then, the vector b_j of basis coefficients of the j -th principal weight function is given by $b_j = \Psi^{-\frac{1}{2}} u_j$, where the vectors u_j are computed as the solutions to the eigenvalue problem

$$n^{-1} \Psi^{\frac{1}{2}} A' A \Psi^{\frac{1}{2}} u_j = \lambda_j u_j,$$

where $n^{-1} \Psi^{\frac{1}{2}} A' A \Psi^{\frac{1}{2}}$ is the sample covariance matrix of $A \Psi^{\frac{1}{2}}$.

The principal components curves estimated by this non-penalized FPCA approach with a B-spline basis can present a lot of variability and have difficult interpretation. To solve this problem two new ways of introducing smoothing in FPCA are proposed in this work. The first one consists of FPCA of the P-spline smoothing of sample curves. The second one is a P-spline version of the smoothed FPCA carried out in Silverman (1996).

2.3 P-spline smoothed functional PCA

In order to control the roughness of the weight functions f_j , the principal components can be computed by maximizing a penalized sample variance that introduces the roughness penalty into the orthonormality constraint. A

continuous penalty based on the integrated squared d-order derivative was first proposed by Silverman (1996). The roughness penalty function is defined by $PEN_d(f) = \int [D^d f(s)]^2 ds$. Following this idea, a penalized sample variance based on a discrete roughness penalty is introduced in this chapter.

Let us consider a B-spline expansion of the weight functions given by $f(t) = \sum_{k=1}^p b_k \phi_k(t) = \phi(t)' b$, where $\phi(t)$ is a B-spline basis, and $b = (b_1, \dots, b_p)$ being the vector of coefficients of the weight function. So, the roughness penalty function exposed before can be written as $PEN_d(f) = b' R_d b$, where R_d is the matrix consisting on the integral of products of d-order derivatives between B-spline basis functions, defined in Equation (1.4). Working with B-splines basis, the continuous roughness penalty function can be approximated by a discrete P-spline roughness penalty function given by $PEN_d(f) = b' P_d b$, where $P_d = (\Delta^d)' \Delta^d$ with Δ^d being the matrix of differences of order d defined in Equation (1.7).

The principal components are now computed as generalized linear combinations of the functional variable that maximize the penalized sample variance defined by

$$\begin{aligned} Pvar \left[\int_T x_i(t) f(t) dt \right] &= \frac{var \left[\int_T x_i(t) f(t) dt \right]}{\|f\|^2 + \lambda PEN_d(f)} \\ &= \frac{b' \Psi V \Psi b}{b' \Psi b + \lambda b' P_d b} = \frac{b' \Psi V \Psi b}{b' (\Psi + \lambda P_d) b}, \end{aligned}$$

whit λ being the smoothing parameter that controls the roughness of the weight function.

The j -th principal component is now defined as in Equation (2.1) and the basis coefficients of the factor loading f_j are obtained by solving

$$\left\{ \begin{array}{l} \max_b \frac{b' \Psi V \Psi b}{b' (\Psi + \lambda P_d) b} \end{array} \right.$$

Let us observe that the first constraint is the usual requirement $\|f\|^2 = 1$ and the second is a modified form of orthogonality that takes into account the roughness of the weight function.

This variance maximization problem can be converted into a eigenvalue problem

$$\Psi V \Psi b = \beta (\Psi + \lambda P_d) b.$$

By applying a factorization (SVD or Choleski) of the form $LL' = \Psi + \lambda P_d$, with L being a lower triangular matrix, then $L^{-1}\Psi V\Psi L^{-1'}(L'b) = \beta(L'b)$ so that the weight coefficients are computed by solving the eigenvalue problem

$$L^{-1}\Psi V\Psi L^{-1'}u = \beta u,$$

with $L'b = u$. This way P-spline smoothed FPCA is reduced to multivariate PCA of the matrix whose rows are the transformed vectors of coefficients $L^{-1}\Psi a_i$. That is multivariate PCA of matrix $A\Psi L^{-1'}$. Finally, the basis coefficients of the principal components curves are given by $b_j = L^{-1'}u_j$ renormalized so that $b_j'\Psi b_j = 1$. Let us observe that in this case the vectors a_i of basis coefficients of the sample curves in terms of the B-spline basis are first estimated by non-penalized least squares approximation that provides $\hat{a}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i'x_i$ with $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$, $i = 1, \dots, n$.

2.3.1 Selection of the smoothing parameter

In order to control the smoothness of the weight function associated to each principal component, selecting a suitable smoothing parameter is very important. In this chapter, leave-one-out cross validation (CV) has been adapted by considering the discrete roughness penalty based on P-splines. It consists of selecting the value of λ that minimizes

$$CV(\lambda) = \frac{1}{p} \sum_{q=1}^p CV_q(\lambda),$$

where

$$CV_q(\lambda) = \frac{1}{n} \sum_{i=1}^n \|x_i - x_i^{q(-i)}\|^2,$$

with $x_i^{q(-i)} = \sum_{l=1}^q \xi_{il}^{(-i)} f_l^{(-i)}$ being the reconstruction of the sample curve x_i in terms of the first q principal components estimated from the sample of size $n - 1$ that includes all sample curves except x_i .

In terms of basis expansions these quadratic distances are given by

$$\begin{aligned}
\|x_i - x_i^{q^{(-i)}}\|^2 &= \int_T \left[x_i(t) - x_i^{q^{(-i)}}(t) \right]^2 dt \\
&= \int_T \left[\sum_{j=1}^p a_{ij} \phi_j(t) - \sum_{l=1}^q \xi_{il}^{(-i)} \sum_{j=1}^p b_{lj}^{(-i)} \phi_j(t) \right]^2 dt \\
&= \int_T \left[\sum_{j=1}^p d_{ij} \phi_j(t) \right]^2 dt \\
&= d_i' \Psi d_i,
\end{aligned}$$

where $d_i = (d_{i1}, \dots, d_{ip})'$ and $d_{ij} = a_{ij} - \sum_{l=1}^q \xi_{il}^{(-i)} b_{lj}^{(-i)}$.

2.4 Functional PCA of P-splines

This version of smoothed FPCA consists of functional PCA of the P-spline smoothing of the original data.

In order to approximate the sample curves observed with error, regression splines are used. As we can see in Chapter 1, their main problem is that they do not control the degree of smoothness. Then, P-spline approach is considered in this section. The number of knots is not so determinant as in regression splines, and can be easily compute by using Ruppert's law of thumb (Ruppert, 2002). In general, the knots of a P-spline must be equally spaced and its number sufficiently large to fit the data and not so large that computation time is unnecessarily high (Eilers and Marx, 2010).

As we can see in the above development there are some important choices related to the P-spline fitting: the smoothing parameter, the order of the penalty, the degree of the B-spline basis and the number of knots. The choice of the smoothing parameter is discussed later in this section. The simplest and most usual choice for the other three parameters that should work well in most applications is use a quadratic penalty, cubic splines and one knot for every four or five observations up to a maximum of about 40 knots (Ruppert, 2002).

The cubic spline functions (piece-wise polynomial curves that has continuous two-order derivatives) tend to be the most popular in applications with smooth functions because of their minimum curvature property (they have

less tendency to oscillate). In fact, cubic spline functions have proven to be a good compromise between accuracy and complexity. Higher order splines allow more strongly curved surfaces to be modeled but also requires more calculations. Usually, degree 3 or 4 is sufficient for B-splines.

Once the P-spline approximations of sample curves have been computed, FPCA is performed on the approximated curves instead of the original sample paths. This way FPCA on P-splines is equivalent to multivariate PCA of the matrix $A\Psi^{\frac{1}{2}}$ where A is the matrix whose rows are the estimated coefficients of the P-spline smoothing of the sample curves in terms of a basis of B-splines.

2.4.1 Selection of the smoothing parameter

As in other smoothing methods, the smoothing parameter of P-splines controls the smoothness of the fitted curve. P-splines penalize distant coefficients so that the larger is the parameter the smoother is the fitted curve. Classical methods for smoothing parameter selection are leave-one-out cross validation (CV), generalized cross validation (Craven and Wahba, 1978), the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). A nonparametric procedure for choosing the P-spline parameters has been performed by Currie and Durban (2002), where mixed model methods based on restricted maximum likelihood (REML) estimation were applied for smoothing parameter selection. The P-spline smoothing of the FPCA (SFPCA) introduced in the next section is not associated with the estimation of a regression model and because of this a direct implementation of the GCV, AIC, BIC and REML criteria is not possible. To make the results comparable with those of the SFPCA approach, CV criterion is used in this chapter to choose the optimum smoothing parameter.

In order to select the same smoothing parameter for the n fitted P-splines, a CV method based on minimizing the mean of the cross validation errors over all P-splines is applied. This CV criterion consists of selecting the smoothing parameter λ that minimizes the expression

$$CVMSE(\lambda) = \frac{1}{n} \sum_{i=1}^n \sqrt{\sum_{k=0}^{m_i} (x_{ik} - \hat{x}_{ik}^{(-k)})^2 / (m_i + 1)},$$

where $\hat{x}_{ik}^{(-k)}$ are the values of the i -th sample path estimated at time t_{ik} avoiding the k -th observation knot in the iterative estimation process.

2.5 Simulation study

The ability of the two smoothed versions of FPCA proposed in this chapter (FPCA of P-splines and P-spline smoothed FPCA) to approximate the true form of the factor loadings functions from a set of smooth curves observed with noise is tested on simulated data. The results are compared with the ones provided by non-penalized FPCA in terms of B-splines and regularized FPCA (RFPCA) based on penalizing the roughness of the principal component curves by its integrated square of the 2-order derivative (see Silverman (1996) for a detailed study).

Let $\{X_t : t \in [0, T]\}$ be the zero mean gaussian process with covariance function $C(t, s) = P \exp(-\alpha|t - s|)$ known as Ornstein-Uhlenbeck process. It can be shown (Van Trees, 1968) that the total variance of this process is T and the principal component weights functions are given by the solutions of the integral eigenequation

$$\int_0^T P \exp(-\alpha|t - s|) f(s) ds = \lambda f(t),$$

whose eigenvalues are

$$\lambda_i = \frac{2P\alpha}{\alpha^2 + b_i^2},$$

with b_i being the positive solutions of

$$\begin{aligned} \tan\left(b_i \frac{T}{2}\right) &= \frac{\alpha}{b_i} \quad (\text{i odd}) \\ \tan\left(b_i \frac{T}{2}\right) &= \frac{-b_i}{\alpha} \quad (\text{i even}). \end{aligned} \quad (2.3)$$

The first fourteen b_i (solutions of Equation (2.3)) can be found in Table 2.1. The eigenfunctions normalized in $[0, T]$ are

$$\begin{aligned} f_i(t) &= \frac{\cos\left(b_i\left(t - \frac{T}{2}\right)\right)}{\left[\frac{T}{2}\left(1 + \frac{\sin(b_i T)}{b_i T}\right)\right]^{\frac{1}{2}}} \quad (\text{i odd}) \\ f_i(t) &= \frac{\sin\left(b_i\left(t - \frac{T}{2}\right)\right)}{\left[\frac{T}{2}\left(1 - \frac{\sin(b_i T)}{b_i T}\right)\right]^{\frac{1}{2}}} \quad (\text{i even}). \end{aligned} \quad (2.4)$$

Table 2.1: The first 14 solutions b_i of the Equation (2.3).

b_1	b_2	b_3	b_4	b_5	b_6	b_7
0.2164	0.8443	1.6020	2.3772	3.1574	3.9397	4.7230
b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}
5.5069	6.2911	7.0757	7.8603	8.6452	9.4301	10.2151

The principal components of the Ornstein-Uhlenbeck process are uncorrelated random variables with normal distribution so that the principal component decomposition of this process is given by

$$X(t) = \sum_{i=1}^{\infty} \lambda_i f_i(t) \xi_i, \quad (2.5)$$

where ξ_i are random variables with distribution $N(0, 1)$. Equation (2.5) truncated in the 14-th term provides a smoothed version of the Ornstein-Uhlenbeck process that explains a 99.4% of its total variance. In order to get noisy observations, a random error $\varepsilon(t)$ with distribution $N(0, \sigma^2)$ was added so that the simulated process was

$$Y(t) = X^{14}(t) + \varepsilon(t). \quad (2.6)$$

The variance of the errors σ^2 was chosen to control the value of $R^2 = \text{var}(X^{14}) / \text{var}(Y)$ close to 0.8. The simulation was made for $T = 4$, $P = 1$ and $\alpha = 0.1$.

In order to test the performance of the four different FPCA approaches compared in this study, 350 samples of 100 sample curves of the contaminated process $Y(t)$ given by Equation (2.6) were simulated at 41 equally spaced knots in the observed domain $[0, 4]$. The first step is to reconstruct the true functional form of the original sample paths $X^{14}(t)$ from the noisy discrete observations $Y(t_k)$ with $t_k = k/10, k = 0, 1, \dots, 40$. Following the methodology proposed in this work, regression splines and P-splines will be computed in terms of B-spline basis. In Figure 2.1 it can be seen an example where an original smooth curve (black solid line), its contaminated curve (black dotted line), the cubic B-spline (blue dashed line) and the quartic B-spline (red dashed dotted line) approximations are superposed for the non-penalized splines (left) and the penalized splines (right) with different number

of basis knots (15, 25 and 30). It can be seen that the reconstructions of the curves provided by cubic and quartic B-splines are very similar for penalized splines independently of the number of knots. In the non-penalized case, the approximation provided by the quartic splines is worse when the number of knots increases by losing control in the extremes. Because of this, a cubic B-splines basis was chosen in this simulation study.

Our main goal is to get an accurate approximation of the true eigenfunctions and the original sample curves. The approximations of the first, second and third eigenfunctions of the Ornstein-Uhlenbeck process by using the four considered FPCA approaches with different number of basis knots (15, 25 and 30) are displayed in Figure 2.2 for one of the 350 simulations. The true eigenfunctions given by the solutions of Equation (2.4) are superposed with their approximations in the same plot. To show the ability of the smoothed FPCA approaches to approximate the original process $X^{14}(t)$, the reconstructions of two different sample paths with the first three PCs approximated by the four considered approaches are displayed in Figure 2.3.

In order to draw general conclusions, the box plots of the MSEs of approximation of the eigenfunctions estimated by using FPCA, FPCA of P-splines, P-spline SFPCA and regularized FPCA with 15, 25 and 30 basis knots for 350 simulations of the Ornstein-Uhlenbeck process were plotted in Figure 2.4. On the other hand, the box plots of the MSEs of the reconstructions of all sample curves with the first three PCs and with all PCs estimated by using the four FPCA approaches for the 350 simulations are displayed in Figures 2.5 and 2.6, respectively.

From the results of this simulation study it can be concluded that the penalized smoothing approaches (FPCA of P-splines, P-spline smoothed FPCA and regularized FPCA) provide approximations of the eigenfunctions and sample curves much more accurate and smoother than the non-penalized FPCA. This is because the regression splines used to estimate the FPCA does not control the degree of smoothness and the roughness and variability of the approximated sample paths increase dramatically with the number of knots. The results provided by the three penalized approaches are quite similar for any number of knots but they get their best behavior when a high number of knots is considered. The accuracy of the approximations provided by the two approaches based on penalizing the roughness of the principal component curves (SFPCA and RFPCA) are very similar and the best approach is FPCA of P-splines because it gives the lowest estimation errors and is

computationally less expensive.

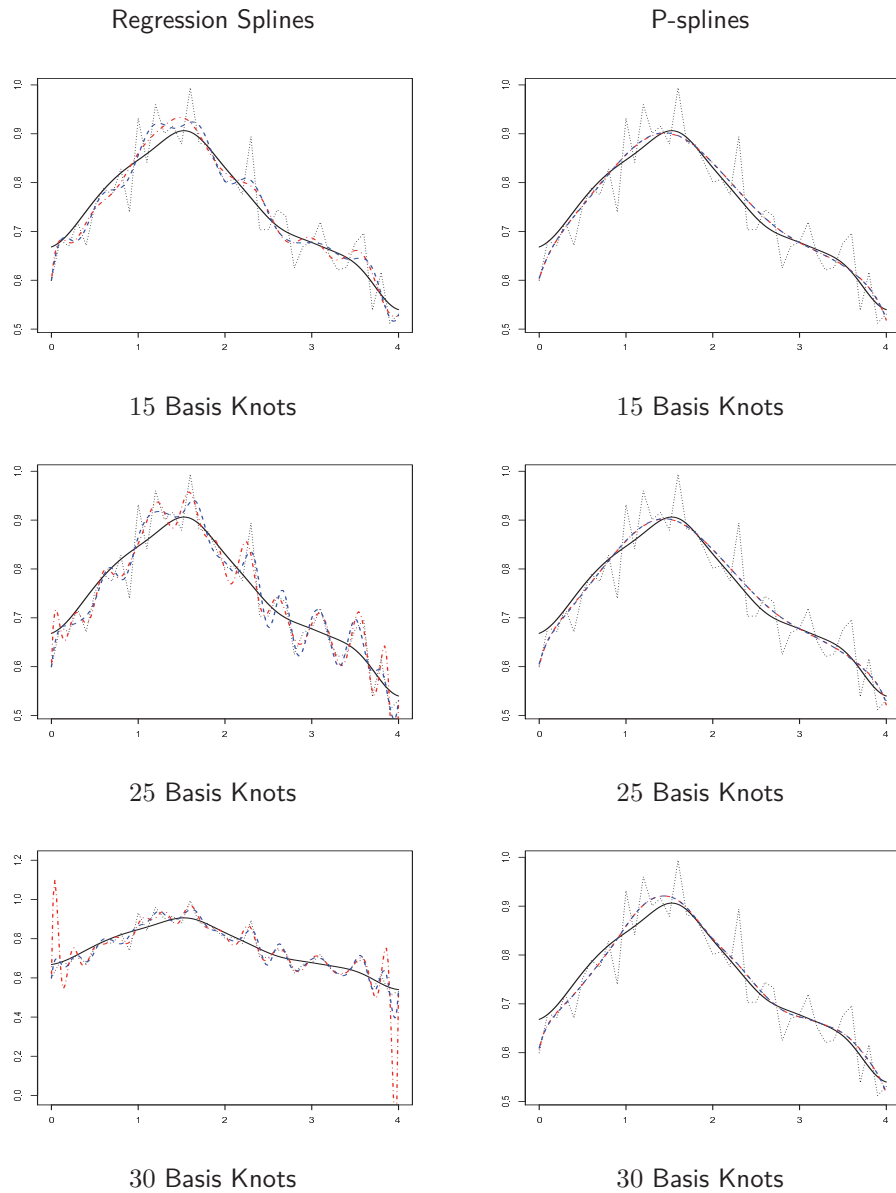


Figure 2.1: Simulation study. Sample path of the Ornstein-Uhlenbeck simulated process approximated by regression splines (left column) and P-splines (right column) using B-splines basis of degree 3 (blue dashed line) and degree 4 (red dashed-dotted line). The original sample paths are displayed in black solid line and the noisy sample paths in black dotted line.

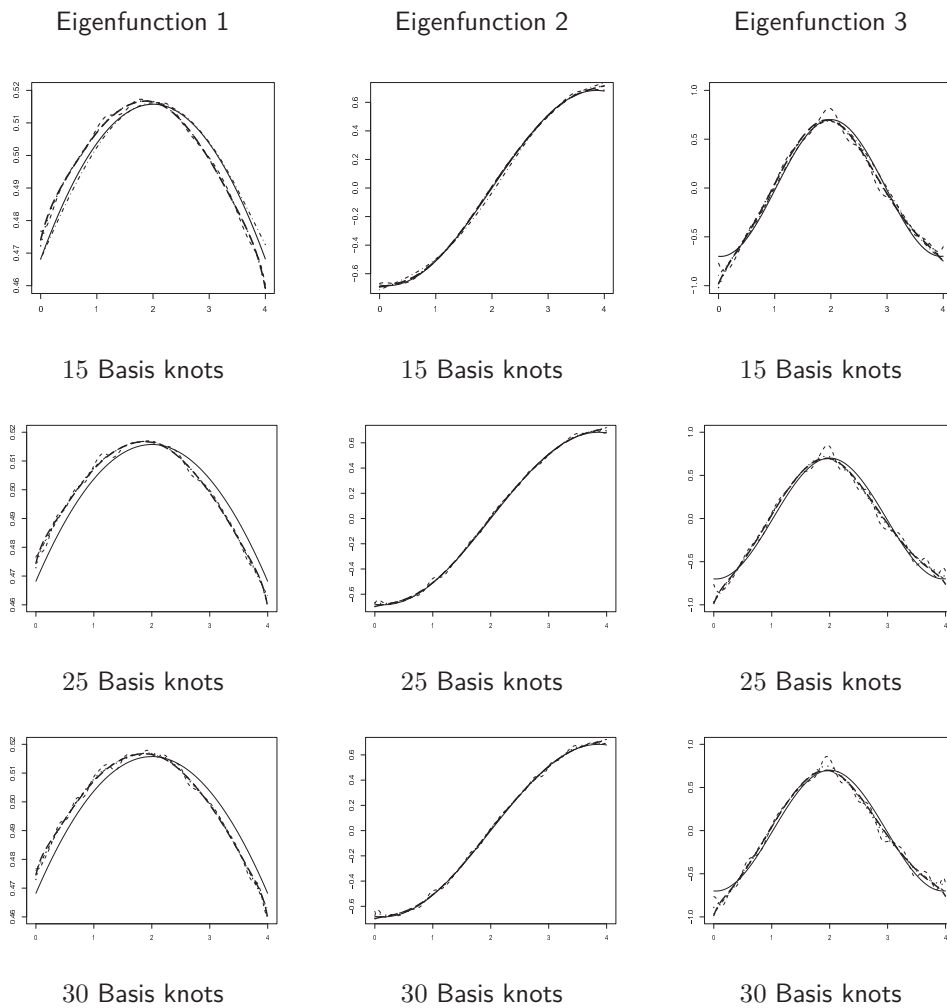


Figure 2.2: Simulation study. First, second and third eigenfunctions of the Ornstein-Uhlenbeck process fitted by FPCA (short dashed line), FPCA via P-splines (long dashed line), SFPCA (dotted line) and RFPCA (dashed-dotted line) for 15 basis knots (λ of P-splines 2.77, λ of SFPCA 0.21 and λ of RFPCA 0.0038), 25 knots (λ of P-splines 13.92, λ of SFPCA 0.31 and λ of RFPCA 0.005) and 30 knots (λ of P-splines 24.37, λ of SFPCA 0.15 and λ of RFPCA 0.005). The true eigenfunctions are represented by black and solid line.

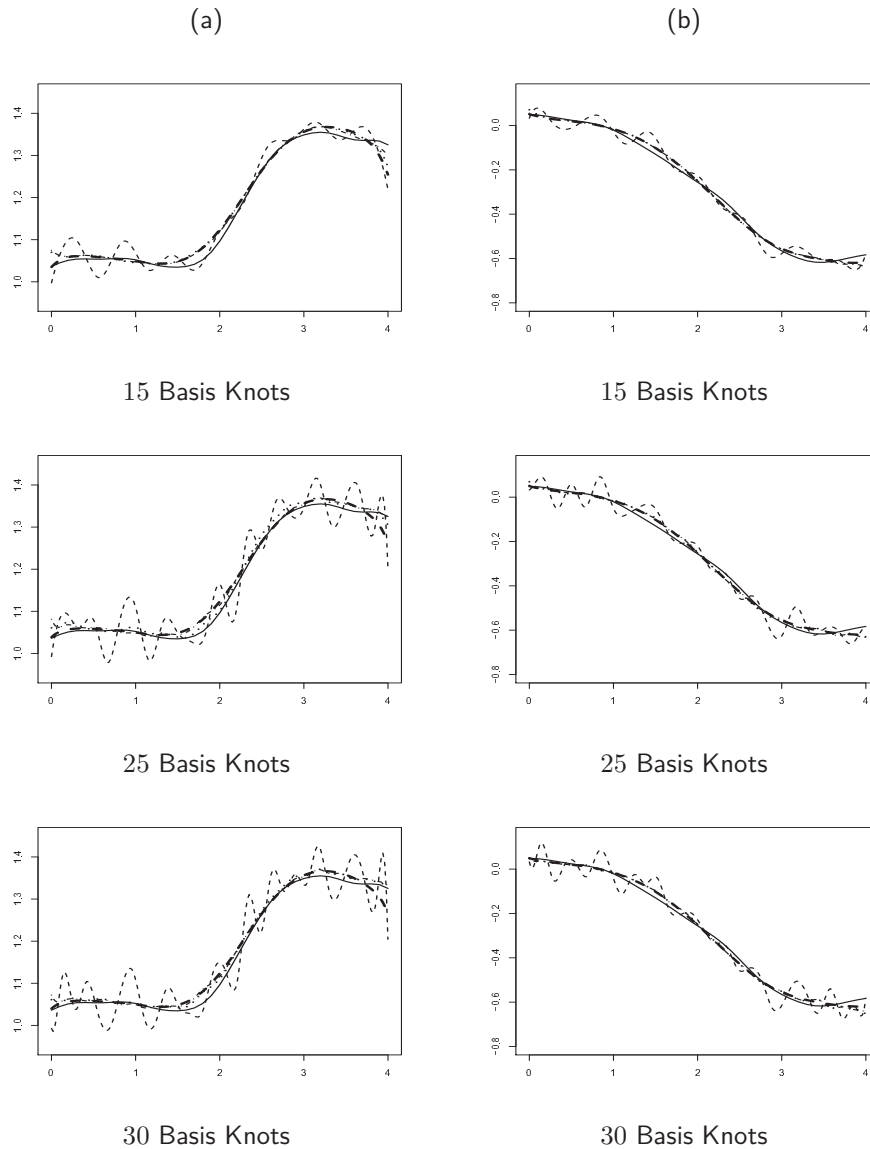


Figure 2.3: Simulation study. Sample path (a) (left column) and sample path (b) (right column) of the Ornstein-Uhlenbeck simulated process reconstructed with the first three PCs estimated by FPCA (short dashed line), FPCA of P-splines (long dashed line), P-spline SFPCA (dotted line) and RFPCA (dashed-dotted line) for 15 basis knots (λ of P-splines 2.46(a) and 2.31(b), λ of SFPCA 0.26(a) and 0.05(b) and λ of RFPCA 0.005(a) and 0.004(b)), 25 knots (λ of P-splines 13.29(a) and 12.66(b), λ of SFPCA 0.10(a) and 0.41(b) and λ of RFPCA 0.005(a) and 0.0035(b)) and 30 knots (λ of P-splines 23.48(a) and 22.59(b), λ of SFPCA 0.92(a) and 1.84(b) and λ of RFPCA 0.0045(a) and 0.005(b)). The true original sample paths are displayed in black and solid line.

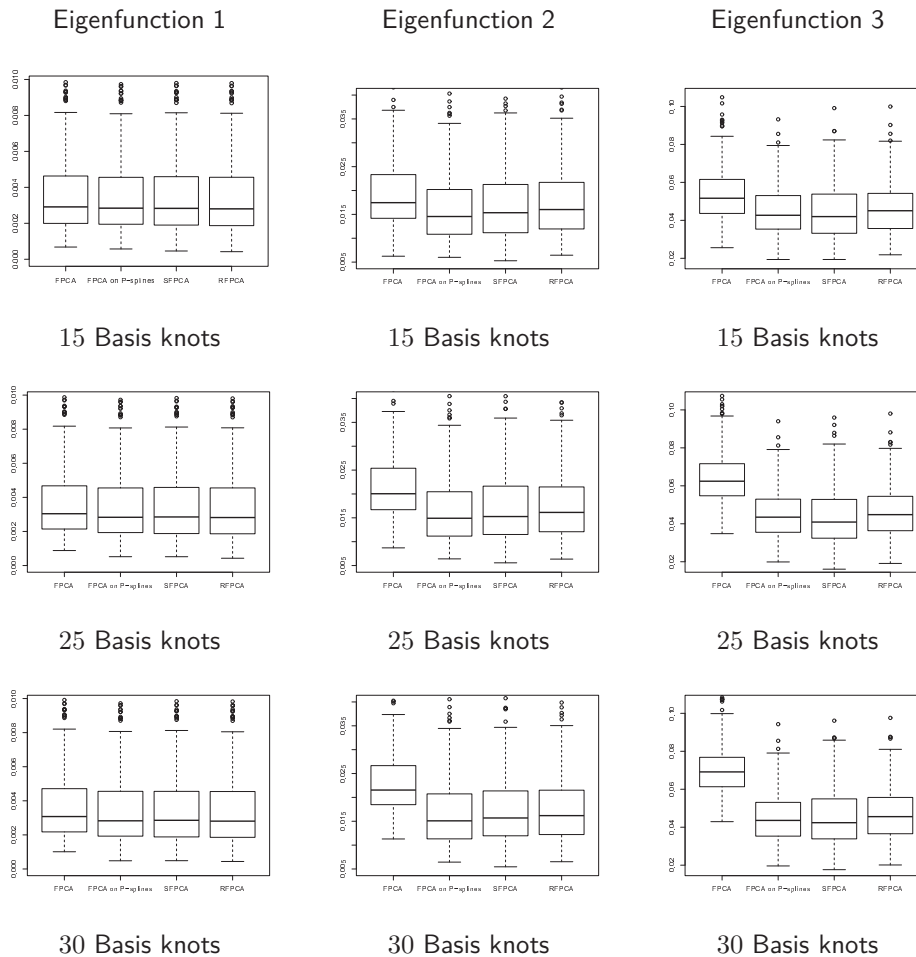


Figure 2.4: Simulation study. Box plots of the MSEs of the eigenfunction 1 (left column), eigenfunction 2 (central column) and eigenfunction 3 (right column) estimated by using FPCA, FPCA of P-splines, P-spline SFPCA and RFPCA with 15, 25 and 30 basis knots on 350 simulations of the Ornstein-Uhlenbeck process.

2.6 Real data application

Finally, we test the performance of the proposed smoothed FPCAs using the Diesel data set downloaded from the website <http://software.eigenvec-tor.com/Data/SWRI/index.html>. This data set was measured by the Southwest Research Institute and consists of a training set of 133 samples and

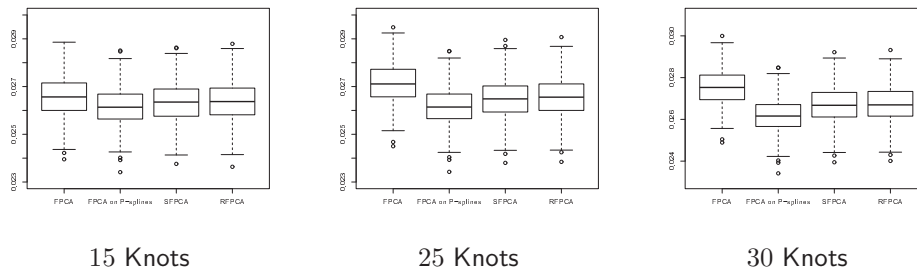


Figure 2.5: Simulation study. Box plots of the MSEs of the reconstructions of sample curves with the first three PCs estimated by FPCA, FPCA of P-splines, P-spline SFPCA and RFPCA with 15, 25 and 30 basis knots for the 350 simulations of the Ornstein-Uhlenbeck process.

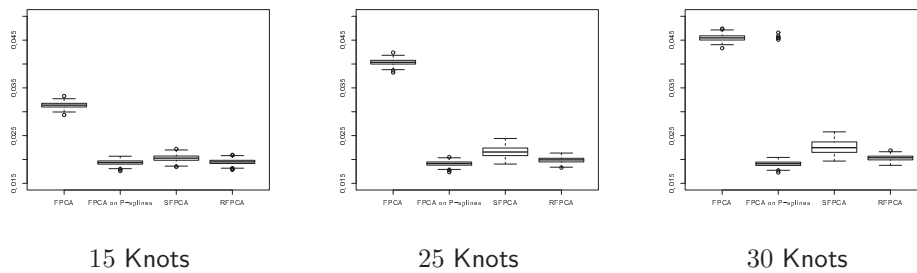


Figure 2.6: Simulation study. Box plots of the MSEs of the reconstructions of sample curves with all PCs estimated by FPCA, FPCA of P-splines, P-spline SFPCA and RFPCA with 15, 25 and 30 basis knots for the 350 simulations of the Ornstein-Uhlenbeck process.

a validation set of 112 samples of the NIR transmission spectra. The NIR spectra of a Diesel sample is measured as a function of wavelengths ranging approximately from 750 to 1550 nm and was used to predict the cetane number of Diesel samples by using functional linear models (Saeys et al., 2008). In this chapter, non-penalized FPCA, FPCA of P-splines and P-spline SFPCA were performed to reduce the dimension of Diesel data and to obtain smooth reconstructions of the NIR transmission spectra. In order to test the good behavior of the smoothed FPCA based on P-spline penalties, cubic B-splines with 30 definition knots were used. The smoothing parameters selected by leave-one-out cross validation are $\lambda = 0.05$ and $\lambda = 0,041$ for FPCA of P-splines and P-spline SFPCA, respectively.

The estimated first and second eigenfunctions are displayed in Figure 2.7. It can be seen that smoothed FPCA approaches provide smoother eigenfunc-

tions than non-penalized FPCA, especially for the second eigenfunction. The smoothest eigenfunctions are provided by FPCA of P-splines. On the other hand, it can be seen in Table 2.2 that the first two PCs estimated by FPCA of P-splines explain a bigger proportion of explained variance than the other two FPCA approaches.

Table 2.2: Application (diesel data). Variances (Var) and percentages of variances (%) explained by the three considered approaches: FPCA, FPCA of P-splines and P-spline SFPCA for Diesel data set.

PC	FPCA		FPCA of P-splines		P-spline SFPCA	
	Var.	%	Var.	%	Var.	%
1	7.07×10^{-4}	85.23	4.44×10^{-4}	87.39	5.73×10^{-4}	86.65
2	6.70×10^{-5}	8.08	4.75×10^{-5}	9.35	5.40×10^{-5}	8.17

In Figure 2.8 two sample paths were reconstructed with the first and second PCs estimated by each one of the FPCA approaches. All sample paths of the test sample were reconstructed with the first and second PCs estimated by using the three FPCA approaches and plotted in Figure 2.9. It is clearly observed that FPCA of P-splines provides the smoothest approximation of sample curves.

2.7 Computational cost

To select the smoothing parameter λ , a cross validation criterion (leave-one-out) is considered in this work. Three smoothing parameters are needed for each simulation, one for computing FPCA of P-splines, other for the P-spline SFPCA and another for regularized FPCA. From the results of the simulation study developed before, it has been concluded that there are not great differences in the approximation errors provided by the three smoothing approaches (FPCA of P-splines, P-spline SFPCA and RFPCA). However, FPCA of P-splines is preferable because its computational cost is lower and the approximation errors are slightly smaller. When we talk about computational cost, we make reference to the CPU time that a computational algorithm consumes during its execution.

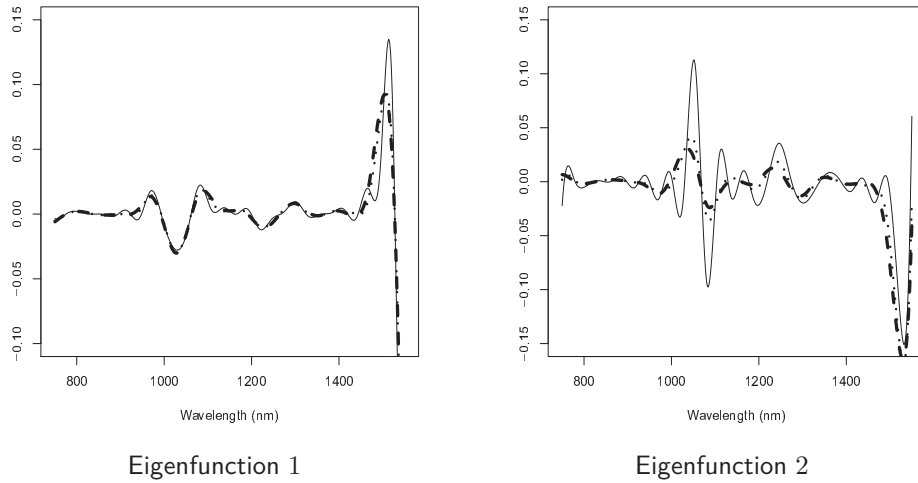


Figure 2.7: Application (diesel data). First (left) and second (right) eigenfunctions estimated by FPCA (solid line), FPCA of P-splines (dashed line) and P-spline SFPCA (dotted line) for Diesel data set.

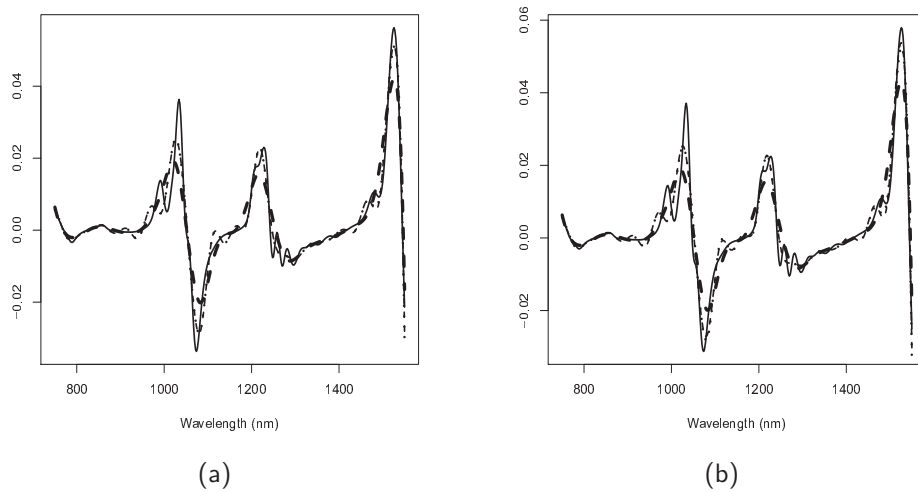


Figure 2.8: Application (diesel data). Noisy sample paths (a) and (b) (solid line) reconstructed by the first and second PCs estimated by FPCA (short dashed line), FPCA of P-splines (long dashed line) and P-spline SFPCA (dotted line).

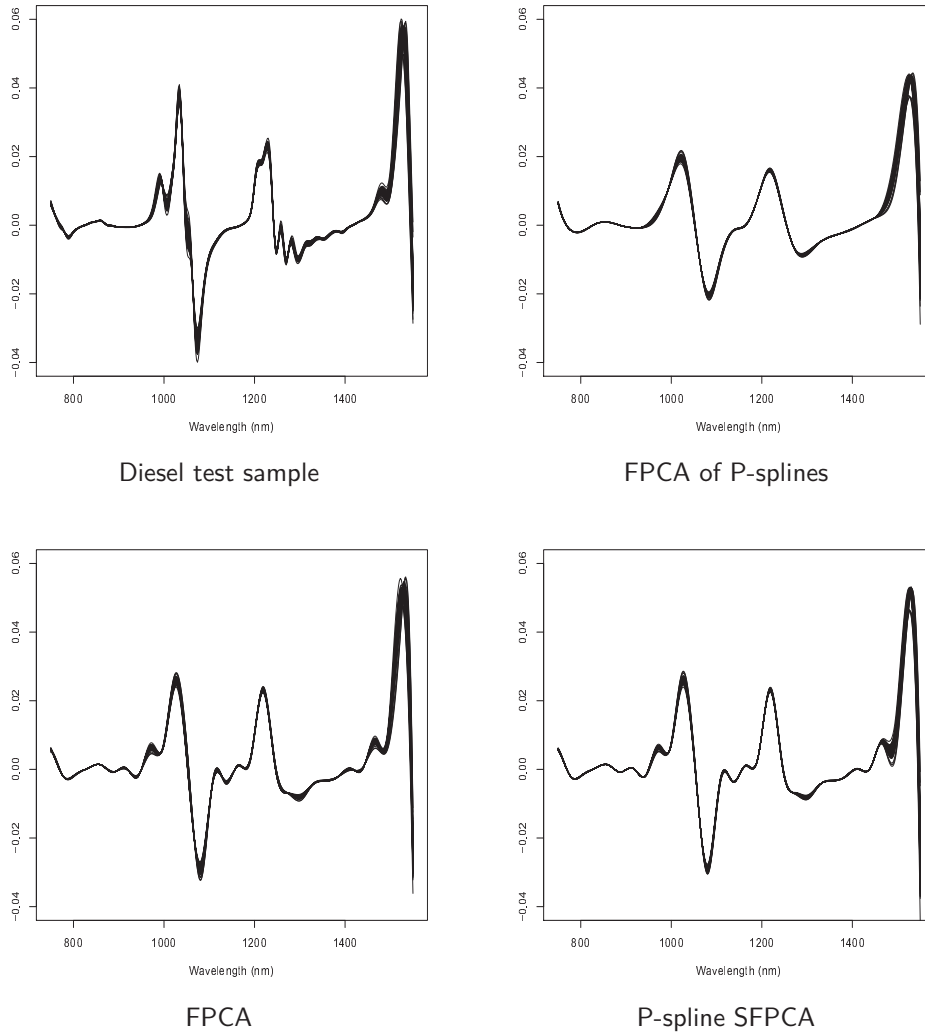


Figure 2.9: Application (diesel data). Original noisy test sample of NIR spectra (top left) reconstructed with the first and second PCs estimated by FPCA (bottom left), FPCA of P-splines (top right), and P-spline SFPCA (bottom right).

In order to compare the computational cost of the cross validation methods used to select λ with FPCA of P-splines, P-spline SFPCA and RFPCA for the 350 simulations of the Ornstein-Uhlenbeck process, the corresponding CPU times were measured by using the function *system.time* of R which returns CPU charged for the execution of user instructions of the calling algorithm. The means of the CPU times required by the different methods are shown in Table 2.3. It is obvious that CPU times spent by CV method with P-spline SFPCA (discrete penalty) and RFPCA (continuous penalty) are much larger than with FPCA of P-splines.

Table 2.3: Mean of the CPU times that the computational algorithms (FPCA of P-splines, P-splines SFPCA and RFPCA) consume during their execution, in the simulation study.

FPCA of P-splines	P-spline SFPCA	RFPCA
104.79	13171.46	11056.90

2.8 Conclusions

Two smoothed FPCA approaches based on P-spline penalties have been proposed in this chapter to control the degree of smoothness of the principal components weight functions estimated from smooth sample curves observed with error. Both approaches are based on B-spline expansion of sample curves and a P-spline penalty that measures the roughness of a function by summing squared d-order differences between adjacent B-spline coefficients. The first smoothed FPCA approach (called FPCA of P-splines) introduces the P-spline penalty in the least squares approximation of the sample curves with B-spline functions (P-splines) and then carries out a non-penalized FPCA on the approximated curves. The second approach approximates the sample curves by non-penalized least squares (regression splines) and then performed a penalized FPCA estimation based on maximizing a penalized sample variance that introduces the P-spline penalty in the orthonormality constraint between principal components.

A simulation study was performed to test the ability of the proposed smoothed FPCA approaches to provide an accurate and smooth estimation of the principal component curves. The results were compared with the estimations provided by non-penalized FPCA of the least squares approximation

of sample curves with B-spline basis and regularized FPCA based on penalizing the roughness of the principal component curves by its integrated squared 2-order derivative. From this simulation study, it can be concluded that the penalized approaches give much more accurate estimations than non-penalized FPCA. This is because FPCA loses control of the smoothness when the dimension of the B-spline base increases. On the other hand, the smoothed FPCA approaches are quite insensitive to the choice of knots so that a relatively large number of equally spaced basis knots is a good election for the definition of the B-spline basis. The advantage of the smoothed FPCA approaches based on P-spline penalties respect to the ones based on penalizing the integrated squared d-order derivatives is that they are mathematically simpler because the difference matrix is easier to compute than the matrix of integrals of products of d-order derivatives between B-spline basis functions (see Bhatti and Bracken (2006) for a detailed study on the calculation of integrals involving B-splines). Finally, it can be concluded that FPCA of P-splines is preferable to P-spline SFPCA and regularized FPCA because its computational cost is lower and the approximation errors are slightly smaller.

Penalized spline approaches for functional logit regression

3.1 Introduction

The aim of the functional logit model (FLoM) is to predict a binary response variable from a functional predictor and also to interpret the relationship between the response and the predictor variables. In the last years, the FLoM was applied in different contexts. A FLoM was applied to predict if human foetal heart rate responds to repeated vibroacoustic stimulation (Ratcliffe et al., 2002). The FLoM was considered in the more general framework of functional generalized linear models in James (2002). A nonparametric estimation procedure of the generalized functional linear model for the case of sparse longitudinal predictors was proposed in Müller (2005). This extension included functional binary regression models for longitudinal data and was illustrated with data on primary biliary cirrhosis. An alternative nonparametric classification method was studied in Ferraty and Vieu (2003).

In order to reduce the infinite dimension of the functional predictor and to solve the multicollinearity problem associated with the estimation of the FLoM, a reduced number of functional principal components can be used as predictor variables to provide an accurate estimation of the functional parameter (Escabias et al., 2004). A climatological application to establish the

relationship between the risk of drought and time evolution of temperatures was carried out by Escabias et al. (2005). The relationship between lupus flares and stress level was analyzed by using a principal component logit model in Aguilera et al. (2008a). A functional PLS based solution was also proposed by Escabias et al. (2007). The problem associated with these approaches is that in many cases the estimated functional parameter is not smooth and therefore difficult to interpret. The main objective of this paper is to solve this problem by introducing different penalties based on P-splines.

The functional linear model was the first regression model extended to the case of functional data. In order to estimate an accurate functional parameter, a smoothing estimation approach based on penalizing the least squares criterion in terms of the squared norm of a B-spline expansion of the functional parameter was introduced by Cardot et al. (2003). A smoothed principal component regression based on ordinary least squares regression on the projection of the covariables on a set of eigenfunctions was also considered. When the functional predictor is corrupted by some error, the functional parameter was estimated by total least squares by using smoothing splines (continuous spline penalty based on the integral of the squared second derivative of the functional parameter) (Cardot et al., 2007). Two versions of functional PCR for scalar response using B-splines and discrete roughness penalty were proposed in Reiss and Ogden (2007). In one of them, the penalty is introduced in the construction of the principal components. In the other one, a penalized likelihood estimation is considered. The smoothing parameter was found by fitting a linear mixed model. These penalized PCR approaches did not consider the functional form of the sample paths but only the approximation in terms of basis functions of the functional parameter. When both the response and the predictor variables are functional, the idea of discrete roughness penalties based on the absolute values of the basis function coefficient differences (corresponding to the LASSO) and the squares of these differences (according to the P-spline methodology) was extended to the functional linear model setting by penalizing the interpretable directions of the regression surface in Harezlak et al. (2007). From a Bayesian point of view, approaches to control the modes of variation in a set of noisy and sparse curves were proposed by Van der Linde (2008) where Demmler-Reinsch basis was used to get smooth weight functions in the functional PCA estimation.

In the general context of functional generalized linear models (FGLM), different penalized likelihood estimations with B-spline basis were proposed

to solve the roughness problem of the functional parameter. The FGLM with P-spline penalty in the log-likelihood criterion was developed in Marx and Eilers (1999). The benefits of this functional model were compared with functional PLS and PCR. A penalized estimation of the functional parameter via penalized log-likelihood was proposed by Cardot and Sarda (2005). This estimation is quite similar to the one provided by Marx and Eilers (1999) with the main difference coming from the continuous penalty that was expressed as the norm of the derivative of given order of the function. A practical mechanism to combine the GLM via penalized log-likelihood, the general additive models (Hastie and Tibshirani, 1990) and the varying-coefficient model (Hastie and Tibshirani, 1993) into a general additive structure was introduced by Eilers and Marx (2002).

In this work, we propose four different methods based on penalized spline (P-spline) estimation of the functional logit regression model by considering the functional form of the sample paths and the functional parameter in terms of B-spline basis expansions. The considered approaches are based on smoothed functional principal component logit regression (FPCLoR) and functional logit regression via penalized log-likelihood.

In the FPCLoR context, three different versions of penalized estimation approaches based on smoothed functional principal component analysis (FPCA) are introduced. On the one hand, FPCA of P-spline approximation of sample curves (Method II) is performed. On the other hand, a discrete P-spline penalty that penalizes the roughness of the principal component weight functions is included in the own formulation of FPCA (Method III). The third smoothed FPCLoR approach is carried out by introducing the penalty in the likelihood estimation of the functional parameter in terms of a reduced set of functional principal components (Method IV). Moreover, direct P-spline likelihood estimation in terms of B-spline functions is also considered (Method V).

The good performance of the proposed methods with respect to non-penalized FPCLoR (Method I) and LDA-FPLS is evaluated via two different data simulations, a functional version of the well known waveform data and a smooth principal component reconstruction of the Ornstein-Uhlenbeck process.

3.2 Functional logit model

The main objective of this chapter consists of estimating the link between a binary random variable Y and a functional predictor $X = \{X(t) : t \in T\}$. It will be assumed without loss of generality that X is a centered second order stochastic process whose sample paths belong to the space $L_2(T)$ of square integrable functions with the usual inner product defined in Equation (1.1). This means that $E(X(t)) = 0, \forall t \in T$.

Let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of the functional variable X and $\{y_i : i = 1, \dots, n\}$ be a random sample of Y associated with 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 π_i is the expectation of Y given $x_i(t)$ modeled as

$$\pi_i = P[Y = 1 | \{x_i(t) : t \in T\}] = \frac{\exp\{\alpha + \int_T x_i(t) \beta(t) dt\}}{1 + \exp\{\alpha + \int_T x_i(t) \beta(t) dt\}}, \quad (3.1)$$

with $i = 1, \dots, n$, α being a real parameter, $\beta(t)$ a functional parameter, and $\{\varepsilon_i : i = 1, \dots, n\}$ independent errors with zero mean. The logit transformations can be expressed as

$$l_i = \ln \left[\frac{\pi_i}{1 - \pi_i} \right] = \alpha + \int_T x_i(t) \beta(t) dt, \quad i = 1, \dots, n. \quad (3.2)$$

In the functional logit model, we have to take into account different aspects. Firstly, we can not continuously observe the functional form of the sample paths. As much we can observe each sample curve $x_i(t)$ in a finite set of discrete sampling points $\{t_{i0}, t_{i1}, \dots, t_{im_i} \in T, i = 1, \dots, n\}$, so that the sample information is given by the vectors $x_i = (x_{i0}, \dots, x_{im_i})'$, with x_{ik} being the observed value for the i -th sample path $x_i(t)$ at time t_{ik} ($k = 0, \dots, m_i$). Secondly, it is impossible to estimate the infinite functional parameter with a finite number of observations n . In order to solve at the same time the two questions, a functional estimation approach based on approximating the sample paths and the functional parameter in terms of basis functions was proposed by Escabias et al. (2007). Different basis such as trigonometric

functions (see Aguilera et al. (1995) and Ratcliffe et al. (2002)), cubic spline functions (see Aguilera et al. (1996) and Escabias et al. (2005)) or wavelet functions (see Ocaña et al. (2008)) can be used depending on the nature of the functional predictor sample paths.

Let us consider that both the sample curves and the functional parameter are approximated as a weighted sum of basis functions as follows:

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad \beta(t) = \sum_{k=1}^p \beta_k \phi_k(t), \quad (3.3)$$

with p being the number of basis functions. Choosing the order of the expansion p is an important problem. If p is increased, the fit to the data is better, but we risk fitting noise or variation that affects the raw data. On the other hand, if p is too small, we may miss some important characteristics of the underlying smooth function.

Then, the FLoM given in Equation (3.2) turns into a multiple logit model whose design matrix is the product between the matrix of basis coefficients of the sample paths and the matrix of inner products between basis functions (Escabias et al., 2004). So, the logit transformations in matrix form are given by

$$L = X\beta, \quad (3.4)$$

where $L = (l_1, \dots, l_n)$ is the vector of logit transformations, $X = (\mathbf{1}|A\Psi)$, with $A = (a_{ij})_{n \times p}$ being the matrix of basis coefficients of the sample paths, $\Psi = (\psi_{jk})_{p \times p}$ the matrix of inner products between basis functions ($\psi_{jk} = \int_T \phi_j(t) \phi_k(t) dt$), $\mathbf{1} = (1, \dots, 1)'$ an n -dimensional vector of ones, and $\beta = (\beta_1, \dots, \beta_p)'$ the vector of basis coefficients of $\beta(t)$.

In order to estimate the multiple logit model given in Equation (3.4) we must first approximate the basis coefficients of each sample curve from its discrete time observations (rows of matrix A). 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 (see Chapter 1 for more details). These approximated sample curves are known as regression splines. The choice of the number of knots is an important problem when working with regression splines because they do not control the degree of smoothness of the estimated curve. This problem is solved in this chapter by using penalized splines. In this case,

the smoothness of the approximated curve is controlled by the smoothing parameter.

3.2.1 Penalized estimation with basis expansions

The log-likelihood function for the multiple model (3.4) is given by

$$\begin{aligned}\mathcal{L}(\beta) &= \sum_{i=1}^n \ln(1 - \pi_i) + \sum_{i=1}^n y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) \\ &= -\sum_{i=1}^n \ln\left(1 + \exp\left(\sum_{j=0}^p X_{ij}\beta_j\right)\right) + \sum_{j=0}^p \left(\sum_{i=1}^n y_i X_{ij}\right) \beta_j.\end{aligned}\quad (3.5)$$

Then, the likelihood equations in matrix form are

$$y'X = \hat{\pi}'X,$$

where $y = (y_1, \dots, y_n)'$, $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_n)'$ is the vector of likelihood estimators of $\pi = (\pi_1, \dots, \pi_n)'$, with $\hat{\pi}_i$ given by

$$\hat{\pi}_i = \frac{\exp\left(\sum_{j=0}^p X_{ij}\hat{\beta}_j\right)}{1 + \exp\left(\sum_{j=0}^p X_{ij}\hat{\beta}_j\right)}$$

and $\hat{\beta}_j$ the likelihood estimators of the basis coefficients of the functional parameter $\beta(t)$ in the FLoM. Solving the likelihood equations by mean of the iterative Newton-Raphson method, the vector of basis coefficients of the functional parameter at iteration t is given by

$$\beta^{(t)} = \beta^{(t-1)} + \left[X' \text{Diag}\left(\pi_i^{(t-1)} (1 - \pi_i^{(t-1)})\right) X\right]^{-1} X' (y - \pi_i^{(t-1)}). \quad (3.6)$$

The maximum likelihood estimate of the parameters of the logit model can be calculated by iterative reweighted least squares as the limit of a sequence of weighted least squares estimates, where the weight matrix changes each cycle. See Agresti (1990) for a detailed study of this least squares procedure.

The estimation of this model is affected by multicollinearity due to the high correlation between the columns of the design matrix. On the one hand, this problem can be solved by logit regression of the response on a set of uncorrelated variables as, for example, principal components. On the other hand, the problem can be solved by using a penalized estimation of the regression

coefficients based on the d -order differences between adjacent coefficients (Le Cessie and Van Houwelingen, 1992). In order to obtain a more accurate and smoother estimation of the functional parameter, this methodology is extended in this section to the functional logit model by introducing a penalty in the log-likelihood estimation of the multiple logit model given by Equation (3.4). This penalty is based on B-spline basis expansions of the sample curves and the functional parameter, and a simple discrete penalty that measures the roughness of the parameter function by summing the squared d -order differences between adjacent B-spline coefficients (P-spline penalty).

Let us consider the basis expansion of the functional parameter given by Equation (3.3). Then, the penalized log-likelihood of the FLoM with logit transformation given by Equation (3.4) is given by

$$\mathcal{L}^*(\lambda, \beta) = \mathcal{L}(\beta) - \frac{\lambda}{2} \beta' P_d \beta,$$

where $\beta = (\beta_1, \dots, \beta_p)'$ is the vector of basis coefficients of $\beta(t)$, λ is the smoothing parameter, and $P_d = (\Delta^d)' \Delta^d$, with Δ^d the matrix of d -order differences defined by Equation (1.7) in Chapter 1.

In this case, the Newton-Raphson solution for the penalized likelihood estimators will be

$$\beta^{(t)} = \beta^{(t-1)} + [X' \text{Diag}(\pi_i^{(t-1)} (1 - \pi_i^{(t-1)})) X + \lambda P_d]^{-1} X' (y - \pi_i^{(t-1)}). \quad (3.7)$$

The number of basis functions p and the smoothing parameter λ are selected by means of a double generalized cross validation (double-GCV) procedure (see Section 3.4.4 for more details). Henceforth, this method will be called Method V.

3.3 Penalized estimation of functional principal component logit regression

As said before, the logit regression model given by Equation (3.4) is affected by multicollinearity. In order to solve the problems of high dimension and high correlation between the covariates of this model, a reduction dimension approach based on using as covariates a reduced set of functional principal components of the predictor curves was proposed (Escabias et al., 2004).

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

$$L = \alpha \mathbf{1} + \Gamma \gamma, \quad (3.8)$$

where $\Gamma = (\xi_{ij})_{n \times p}$ is a matrix of functional principal components of the sample paths $\{x_i(t) : i = 1, \dots, n\}$, γ is the vector of coefficients of the model and α is the intercept.

An accurate estimation of the functional parameter can be obtained by considering only a set of q optimum principal components as predictor variables, so that $\Gamma = (\xi_{ij})_{n \times q}$ ($q < p$). Then, the vector β of basis coefficients is given by $\beta_{p \times 1} = F_{p \times q} \gamma_{q \times 1}$, where the way of estimating F depends on the kind of functional principal component analysis (FPCA) used to estimate the functional model and the kind of likelihood estimation (penalized or non-penalized). According to it, four different methods are considered in this chapter.

3.3.1 Method I: non-penalized FPCLoR

A simple way to estimate the functional parameter is by means of non-penalized functional logit regression on an optimum set of principal components. This method known as non-penalized functional principal component logit regression (FPCLoR) was performed by Escabias et al. (2004).

In practice, functional PCA is estimated from discrete time observations of each sample curve $x_i(t)$ which is approximated in terms of basis functions. If we assume that the sample curves are represented in terms of basis functions as in expression (3.3) the functional PCA is then equivalent to the multivariate PCA of $A\Psi^{\frac{1}{2}}$ matrix, with $\Psi^{\frac{1}{2}}$ being the square 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 estimated in terms of principal components is given by $F = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times q}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$. In this case, the matrix of basis coefficients A is computed by using least squares approximation with B-spline basis and γ is estimated by maximum likelihood without penalty. The optimum number of principal components of the predictor curves used as covariates is chosen by GCV (see Subsection 3.4.3 for more details).

3.3.2 Method II: FPCLoR on P-spline smoothing of the sample curves

When the sample paths are observed with noise, the estimation of the FLoM based on FPCA of regression splines provides a noisy functional parameter. This is because of regression splines do not control the smoothness of the sample paths. In order to smooth the sample curves, P-spline approximation is carried out. Therefore, a penalized estimation of the FLoM based on FPCA of the P-spline approximation of the sample curves is proposed. In this case, the smoothing parameter of P-splines is chosen by leave-one-out cross validation (Section 3.4.1).

Once the P-spline approximation of sample curves has been performed, the multivariate PCA of $A\Psi^{\frac{1}{2}}$ matrix is carried out as explained above, being A the basis coefficients matrix estimated with P-spline penalty. The difference between smoothed FPCA via P-splines and non-penalized FPCA is only the way of computing the basis coefficients (rows of matrix A), with or without penalty, respectively. Then, an optimum set of principal components is selected and the FPCLoR is carried out. In this case, $F = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times q}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$. In this method, γ is estimated via maximum likelihood without penalty.

The optimum number of principal components is chosen by GCV (see Subsection 3.4.3 for more details).

3.3.3 Method III: FPCLoR on P-spline smoothing of the principal components

In this section, we propose obtaining the principal components by maximizing a penalized sample variance that introduces a discrete penalty in the orthonormality constraint between weight principal component functions.

Taking into account the basis expansion of the sample paths given by (3.3), the principal component weight function f_j admits the basis expansion $f_j(t) = \sum_{k=1}^p b_{jk} \phi_k(t)$. In Chapter 2, we can see that P-spline smoothing of the functional principal components analysis consists of a classical PCA of the matrix $A\Psi(L^{-1})'$, where L comes from the Cholesky decomposition $LL' = \Psi + \lambda P_d$, with λ being the smoothing parameter estimated by leave-one-out cross validation, Ψ the matrix of inner products between basis func-

tions, and P_d the discrete penalty matrix defined in Chapter 1.

Once the P-spline smoothing of the FPCA is computed, we carry out the FPCLoR on an optimum set of principal components obtained by the P-spline smoothing of FPCA. Then, the estimated vector β 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 $A\Psi(L^{-1})'$ and γ is estimated by the maximum likelihood criterion without penalty. The optimum number of principal components to be included in the model as regressors is chosen by GCV (see Subsection 3.4.3).

3.3.4 Method IV: FPCLoR with P-spline penalty in the maximum likelihood estimation

As developed in Reiss and Ogden (2007) for the functional linear model, we propose a smoothed version of FPCLoR that uses B-splines and roughness penalty in the regression. This penalized regression version of FPCLoR incorporates a penalty in the maximum likelihood estimation.

Taking into account the FLoM in terms of non-penalized principal components, and the Equation (3.3), the estimator of the basis coefficients of the functional parameter corresponds to $\hat{\beta} = F\hat{\gamma}$, where F is exactly the same as in Section 3.3.1 and γ is estimated by means of penalized likelihood.

Now the design matrix corresponds to $X = (\mathbf{1}|\Gamma)$, where $\Gamma = (\xi_{ij})_{n \times q}$ is a matrix of an optimum set of q functional principal components of the sample paths. Then, the penalized log-likelihood of the functional principal components logit model (3.4) is given by

$$\mathcal{L}^*(\lambda, \gamma) = \mathcal{L}(\gamma) - \frac{\lambda}{2} \gamma' P_d \gamma,$$

with $\gamma = (\gamma_1, \dots, \gamma_q)'$ being the vector of the regression coefficients, P_d the discrete penalty matrix defined in Chapter 1, with dimension $(q \times q)$ in this case, and $\mathcal{L}(\gamma)$ given by Equation (3.5).

The optimal number of principal components and the smoothing parameter are chosen by a double GCV procedure (see Subsection 3.4.4 for more details).

3.4 Model Selection

Penalized FPCLoR requires selecting an optimal number q of functional principal components and the smoothing parameter λ . Using P-spline smoothing of FPCA, the problems of high dimension, multicollinearity, and roughness in the covariables are solved. As cited in Reiss and Ogden (2007), and according to Marx and Eilers (1999) and Cardot et al. (2003), it is often assumed that the number of basis functions considered for computing P-splines has little impact as long as there are many knots to capture the variation in the functional parameter. Methods based on smoothed FPCA (Methods II and III) select λ in a previous step to the selection of the number q of principal components.

On the other hand, when the smoothing is applied in the likelihood estimation of the functional parameter coefficients (Methods IV and V), the sample paths are approximated by regression splines. It is known that regression splines do not control the degree of smoothness in the curves. Therefore, the selection of the number of predictor variables (non-penalized principal components for Method IV and basis functions for Method V) is essential. The optimal number of predictors and the smoothing parameter are selected in these cases by a double-GCV procedure.

3.4.1 Choosing λ in Method II

In Method II (Section 3.3.2) the smoothing parameter λ was selected prior to the regression. In order to select the same smoothing parameter for the n fitted P-splines, a leave-one-out cross validation (CV) method based on minimizing the mean of the cross validation errors over all P-splines is applied in this chapter. This CV criterion consists of selecting the smoothing parameter λ that minimizes the expression

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \sqrt{\sum_{k=0}^{m_i} (x_{ik} - \hat{x}_{ik}^{(-k)})^2 / (m_i + 1)},$$

where $\hat{x}_{ik}^{(-k)}$ are the values of the i -th sample path estimated at the time t_{ik} avoiding the k -th observation knot in the iterative estimation process. The number of observation knots of the i -th sample path corresponds to $m_i + 1$.

3.4.2 Choosing λ in Method III

As in the previous section, selecting a suitable smoothing parameter is very important to control the smoothness of the weight function associated with each principal component. In this chapter, CV (leave-one-out) method described in Ramsay and Silverman (2005) has been adapted by considering the discrete roughness penalty based on P-splines. It consists of selecting the value of λ that minimizes

$$CV(\lambda) = \frac{1}{p} \sum_{q=1}^p CV_q(\lambda),$$

where

$$CV_q(\lambda) = \frac{1}{n} \sum_{i=1}^n \|x_i - x_i^{q(-i)}\|^2,$$

with

$$x_i^{q(-i)} = \sum_{\ell=1}^q \xi_{i\ell}^{(-i)} f_{\ell}^{(-i)}$$

being the reconstruction of the sample curve x_i in terms of the first q principal components estimated from the sample of size $n - 1$ that includes all sample curves except x_i .

3.4.3 Choosing the number of principal components in Methods I, II, and III

The optimal number q of functional principal components for Methods I, II, and III is chosen by the GCV procedure following the notes given in Craven and Wahba (1978) and Ramsay and Silverman (2005). The objective is to minimize

$$GCV(q) = \left(\frac{n}{n - \text{tr}(H^q)} \right) \left(\frac{MSE(q)}{n - \text{tr}(H^q)} \right), \quad (3.9)$$

where

$$MSE(q) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^q)^2$$

and H^q is the "hat" matrix given by

$$H^q = W_q^{1/2} X (X' W_q X)^{-1} X' W_q^{1/2},$$

with $W_q = \text{Diag}[\hat{\pi}_i^q (1 - \hat{\pi}_i^q)]$ as the weight matrix. The design matrix X depends on the considered method as follows:

Method I: $X = (1|\Gamma)$, with Γ being the matrix comprising the columns of the first q functional principal components of $A\Psi^{1/2}$, with A the matrix of basis coefficients of the sample paths estimated via regression splines, and $\Psi^{1/2}$ the square root of the matrix of the inner products between B-spline basis functions.

Method II: $X = (1|\Gamma)$, with Γ being the matrix comprising the columns of the first q functional principal components of $A\Psi^{1/2}$, with A the basis coefficients estimated via penalized splines (P-splines).

Method III: $X = (1|\Gamma)$, with Γ being the matrix comprising the columns of the first q functional principal components of $A\Psi(L^{-1})'$, with A the matrix of basis coefficients of the sample paths estimated via regression splines, and L given by the Cholesky decomposition.

3.4.4 Choosing the number of predictors and the smoothing parameter in Methods IV and V

In Methods IV and V the log-likelihood is penalized and the parameters of the model are simultaneously chosen by a double-GCV. In Method IV, the double-GCV consists in computing the GCV error (given by expression (3.9)) for each number of principal components q and each λ of a grid of possible values. Then, q is selected by minimizing the mean of the GCV error over all possible values of λ . Once q is selected, the value of λ with the lowest GCV error is chosen. In Method V, the procedure is the same by replacing the number q of principal components by the number p of basis functions.

The design matrix X of Methods IV and V corresponds to

Method IV: $X = (1|\Gamma)$, with Γ being the matrix comprising the columns of the q first functional principal components of $A\Psi^{1/2}$ and A the matrix of basis coefficients of the sample paths estimated via regression splines.

Method V: $X = (1|A\Psi)$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines.

When the log-likelihood criterion is penalized by using P-splines, the "hat" matrix is given by

$$H = W^{1/2} X (X' W X + \lambda P_d)^{-1} X' W^{1/2},$$

with P_d the discrete penalty matrix defined in Chapter 1.

3.5 Simulation study

The good performance of the proposed penalized estimation approaches to estimate the parameter function and to predict the response is evaluated in this section on two different simulation schemes, and the results compared with the ones provided by non-penalized FPCLoR (Method I).

On the other hand, the ability of the proposed approaches to forecast a binary response and classify a set of curves has also been compared with a competitive classification procedure as the partial least squares approach for functional linear discriminant analysis (FLDA-PLS) introduced by Preda et al. (2007) and its basis expansion estimation with B-spline basis proposed in Aguilera et al. (2010b). It is important to clarify that we can compare our results with the prediction errors and classification rates given by this procedure but the estimated parameter functions are not comparable because they correspond to different regression models from a theorist point of view.

3.5.1 Case I: simulation of waveform data

This data set was introduced by Breiman et al. (1984) and used later by Hastie et al. (1994), Ferraty and Vieu (2003), and Escabias et al. (2007). Following the simulation scheme developed in Escabias et al. (2007), 1000 curves of two different classes of sample curves were simulated with 500 curves for each one according to the random functions

$$x(t) = u h_1(t) + (1 - u) h_2(t) + \varepsilon(t) \quad (\text{class 1}),$$

$$x(t) = u h_1(t) + (1 - u) h_3(t) + \varepsilon(t) \quad (\text{class 2}),$$

with u and $\varepsilon(t)$ being uniform and standard normal simulated random variables, respectively, and

$$h_1(t) = \max\{6 - |t - 11|, 0\}, \quad h_2(t) = h_1(t - 4), \quad h_3(t) = h_1(t + 4).$$

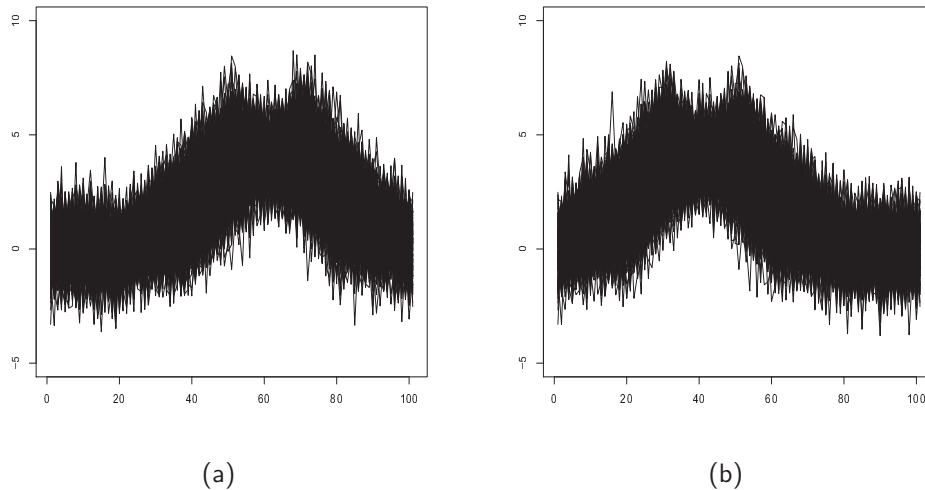


Figure 3.1: Simulation study. Case I. Simulated sample curves for class 1 (a) and class 2 (b) in one of the 100 simulations.

Each sample curve was simulated at 101 equally spaced points in the interval $[1, 21]$.

An example of simulated sample paths for class 1 (a) and class 2 (b) is shown in Fig. 3.1. The binary response variable was defined as $Y = 0$ for the curves of the first class and $Y = 1$ for the ones of the second class. After simulating the data, least squares approximation (with and without penalty) in terms of the cubic B-spline functions defined on 30 equally spaced knots in the interval $[1, 21]$ was performed for each sample curve. When working with P-splines, the number of basis knots is not so critical and only a large number of equally spaced knots is needed. The choice of the P-splines parameters was discussed by Eilers and Marx (1996), Ruppert (2002), and Currie and Durban (2002). For the case of equally spaced observations, they conclude that using one knot for every four or five observations up to a maximum of 40 knots is often sufficient.

In order to corroborate the good performance of the penalized estimation approaches proposed in this chapter, 100 repetitions of this simulation scheme were carried out. The functional parameter estimated by means of the five different methods presented in previous sections are displayed in Figure 3.2 for

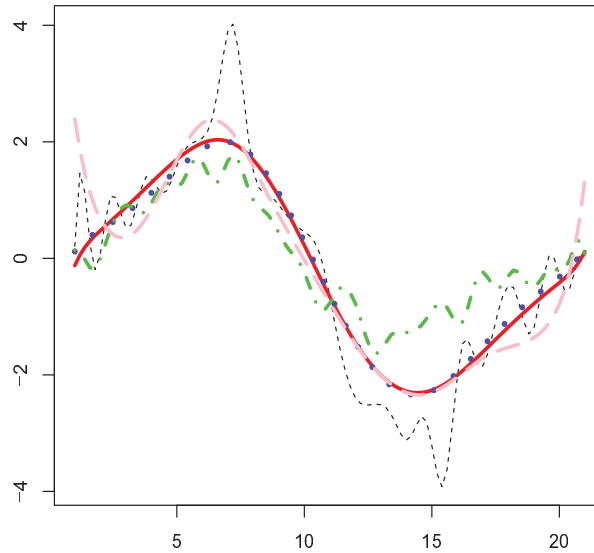


Figure 3.2: Simulation study. Case I. Estimated functional parameter for one of the 100 simulations. The functional parameter is estimated by Method I (black short dashed line), Method II (red solid line), Method III (blue dotted line), Method IV (green dashed and dotted line), and Method V (pink large dashed line).

one of the simulations. The mean of the estimated functional parameters over the 100 simulations is plotted in Figure 3.3 for each of the five estimation approaches and for FLDA-PLS next to confidence bands computed as the mean ± 2 the standard deviation. The functional parameter estimated by FLDA-PLS (discriminant function) is not comparable to the others because is associated with different regression models.

Let us observe that there are important differences between the estimations provided by the non-penalized FPCLoR approach (Method I) and the other four methods based on penalized estimation of the FLoM. The functional parameter estimated by non-penalized FPCLoR (Method I) is not smooth and affected by high variability. It is therefore difficult to interpret and needs to be smoothed. The estimations provided by Methods II, III, and V are quite simi-

Table 3.1: Simulation study. Case I. Mean and standard deviation (S.D.) for the GCV errors of the models estimated by Methods I, II, III, IV, and V.

Method	GCV error	
	Mean	S.D
Method I	0.00003	0.000008
Method II	0.00003	0.000007
Method III	0.00002	0.000007
Method IV	0.00014	0.000112
Method V	0.00022	0.000120

lar, but sometimes the estimations provided by Method V are over-smoothed and lose the control in the extremes of the observation interval. On the other hand, when the P-spline penalty is introduced in the log-likelihood criterion of a FPCLoR model (Method IV), the estimated functional parameter is smoother than the one given by Method I but it is not smooth enough and is affected by some variability. Therefore, the necessity of using smoothed functional principal components as explicative variables is obvious. The best estimations are achieved with Methods II and III, providing the smoothest parameter functions with the least variability.

In order to compare the goodness of fit and the forecasting ability of the five estimation approaches the box-plots related to the area under ROC curve and the MSE distributions (on 100 test samples) are shown in Figure 3.4. It can be observed that the Methods I, II, and III based on non-penalized principal component logit regression result in much more accurate predictions than Methods IV and V based on penalized likelihood estimation. Among them, Method II achieves the highest area under ROC curve and Method III the smallest MSE and GCV error. Let us observe that the FLDA-PLS approach gets the highest prediction error, but has good classification ability similar to Methods II and III.

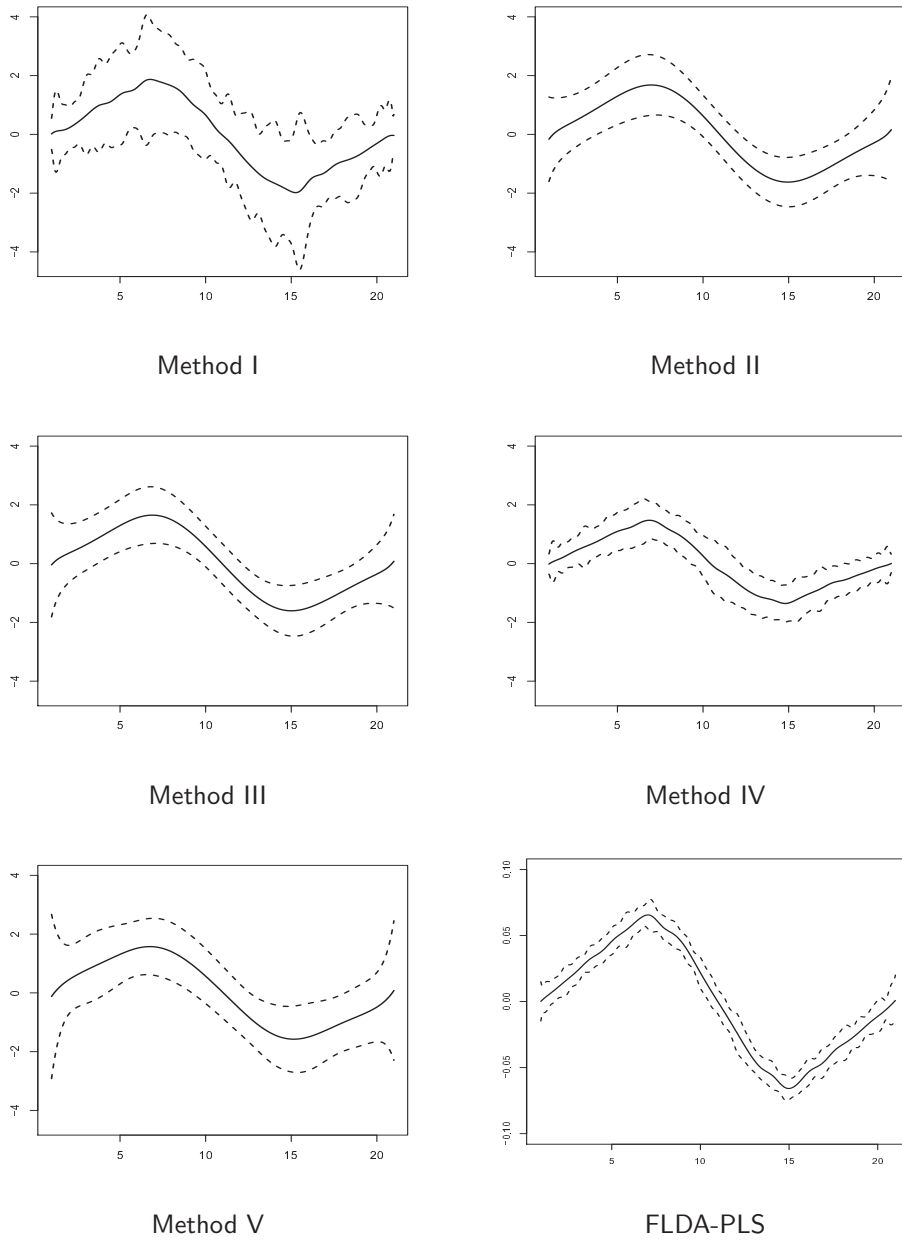


Figure 3.3: Case I. Mean of the functional parameters and the confidence bands (computed as the mean ± 2 the standard deviation) estimated by Methods I, II, III, IV, V, and FLDA-PLS over the 100 simulations.

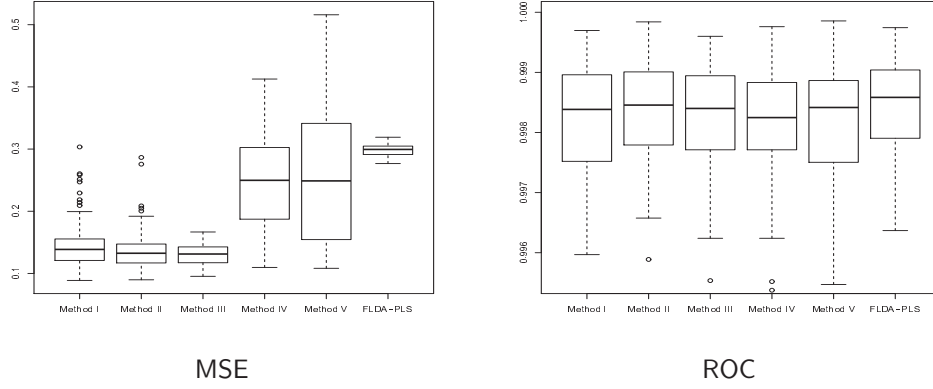


Figure 3.4: Simulation study. Case I. Area under ROC curve and MSE distribution (for the test samples of the 100 repetitions) given by Methods I, II, III, IV, V, and FLDA-PLS.

3.5.2 Case II: simulation of the Ornstein-Uhlenbeck process

In order to obtain more general conclusions about the behavior of the proposed methods a second simulation study where the functional parameter is known has been developed.

Let us consider $\{O_t : t \in [0, T]\}$ the well known zero mean gaussian process known as Ornstein-Uhlenbeck process. The simulated sample paths were computed taking into account the decomposition of this process in terms of principal components truncated at the 14th term

$$O^{14}(t) = \sum_{i=1}^{14} \lambda_i f_i(t) \xi_i,$$

with λ_i and f_i being the eigenvalues and eigenfunctions associated with the covariance function given by $C(t, s) = P \exp(-\alpha|t - s|)$, and ξ_i being the corresponding principal components that have distribution $N(0, 1)$. This principal component reconstruction is a smooth version of the Ornstein-Uhlenbeck process that explains 99.4% of its total variance.

In order to have noisy observations, a random error $\varepsilon(t)$ with distribution $N(0, \sigma^2)$ was added so that the simulated process is given by

$$X(t) = O^{14}(t) + \varepsilon(t). \quad (3.10)$$

The variance of the errors σ^2 was chosen by controlling $R^2 = Var[O^{14}]/Var[X]$ close to 0.8. The parameters used for the simulation were $T = 4$, $P = 1$, and $\alpha = 0.1$. In this study, 200 samples of 100 and 50 sample curves of the contaminated process given by Equation (3.10) were simulated for training and test samples, respectively, at 41 equally spaced knots in the interval $[0, 4]$.

In order to simulate the binary response associated with each sample path x_i , we have considered the parameter function

$$\beta(t) = 6 \cos(0.25\pi t) - 0.5 \sin(0.25\pi t)$$

and computed the expectations π_i according to Equation (3.1). Then, the associated response value y_i was simulated by a Bernoulli distribution with parameter π_i .

Let us remember that the main purpose of this work is to improve the estimation of the functional parameter in functional logit regression, providing in addition a good classification rate. In order to check the ability of the proposed penalized spline approaches to estimate the functional parameter of the logit model provided by the six methods, the mean of the estimated functional parameters over the 200 simulations is plotted in Figure 3.5 next to the original parameter function and the confidence bands (computed as the mean ± 2 the standard deviation). The integrated mean squared error with respect to the original functional parameter was also computed for each method by using the following expression:

$$IMSE\beta = \left(\frac{1}{T} \int_T (\beta(t) - \hat{\beta}(t))^2 \right)^{1/2}.$$

The box plots with the distribution of the $IMSE\beta$ for the five estimation approaches of the functional parameter associated to the logit model are displayed in Figure 3.6. The means and standard deviations of these errors appear in Table 3.2.

Let us observe that Methods I (non-penalized FPCLoR approach) and IV provide the least smooth estimates with the worst results given by Method IV that is affected by high variability. On the other hand, Methods II and III provide again similar results with smoother estimates affected by high variability in the extremes of the observation interval. By observing the estimated mean functions it can be observed again that the estimations provided by Method V are over-smoothed and have less variability than the one given by Methods II and III. The integrated errors with respect to the original parameter function

Table 3.2: Simulation study. Case II. Mean and standard deviation of the $IMSE_{\beta}$ distribution.

	Method I	Method II	Method III	Method IV	Method V
Mean	3.1893	1.8931	1.8166	8.0691	2.5332
SD	9.4394	2.0289	1.8645	4.1827	1.3898

are also higher for Method V than for Methods II and III. The discriminant function associated with the FLDA-PLS approach is noisy and affected of a very high degree of variability (see Figure 3.7).

The forecasting performance and classification ability of the six methods can be tested by comparing the distributions of the mean squared error (MSE) and ROC area displayed in Figure 3.8. According to the MSE, Methods III and V are quite similar, providing the smallest prediction errors, while FLDA-PLS gives the highest prediction errors. With respect to the ROC area, Methods III and V achieve also the highest values followed by Method II and FLDA-PLS. On the other hand, Method IV provides the worst classification performance (smallest area under the ROC curve), although in all cases the ability of the considered methods to classify the curves is very good with a median greater than 93%. From this simulation, we can conclude that Method III provides an accurate estimation of the functional parameter and has the best classification ability followed by Methods V and II that give similar results. In addition, Method III outperforms competitive methods such as FLDA-PLS in both predictive and classification ability.

3.6 Conclusions

In order to solve the problem of multicollinearity in functional logit regression and to control de smoothness of the functional parameter estimated from noisy smooth sample curves, four different penalized spline (P-spline) estimations of the functional logit model are proposed in this chapter. Let us take into account that the aim of logit model is not only to classify a set of curves in two groups but mainly to interpret the relationship between the binary response and the functional predictor in terms of the functional parameter. Because of this, our main purpose is to improve the estimation of the functional parameter of a functional logit model, providing in addition a good classification rate.

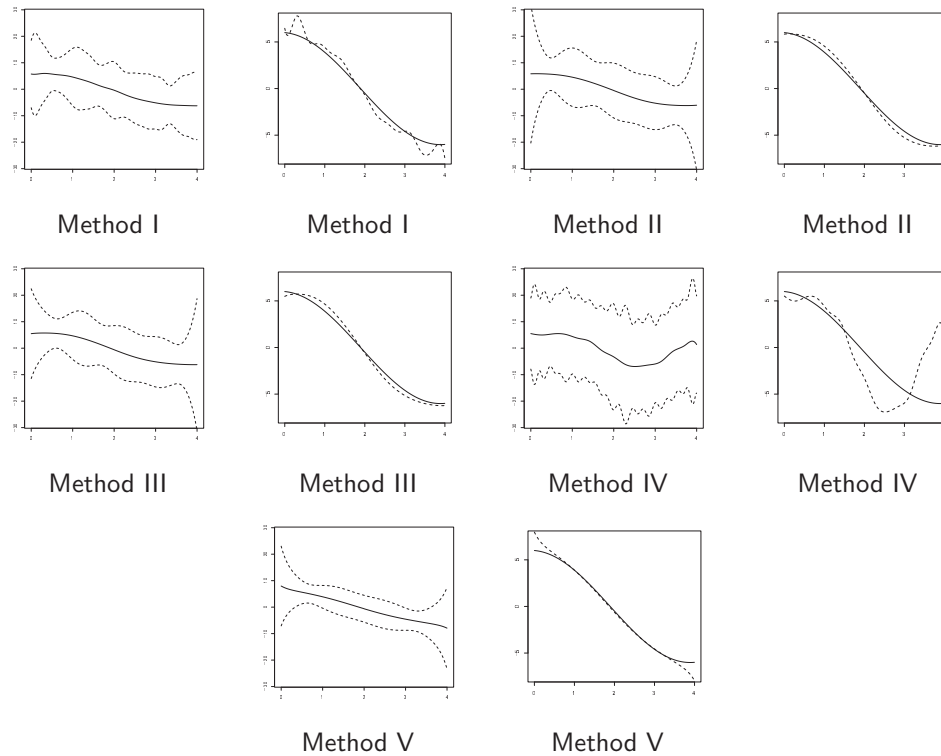


Figure 3.5: Simulation study. Case II. Mean of the functional parameters and the confidence bands (computed as the mean ± 2 the standard deviation) (at left) and the true functional parameter (solid line) superposed with the mean of the estimated functional parameters (dashed line) (at right) provided by Methods I, II, III, IV, and V over 200 simulations.

A P-spline penalty measures the roughness of a curve in terms of differences of order d between coefficients of adjacent B-spline basis functions. The proposed smoothing approaches are based on B-spline expansion of the sample curves and the parameter function, and P-spline estimation of the functional parameter. The difference is in how to introduce the penalty in the model. Three of the considered approaches (Methods II, III and IV) are based on functional principal component logit regression that consists in regressing the binary response on a reduced set of functional principal components. In Method II the P-spline penalty is introduced by performing the functional PCA on the P-spline least squares approximation of the sample curves from discrete observations. Method III introduces the P-spline penalty in the own

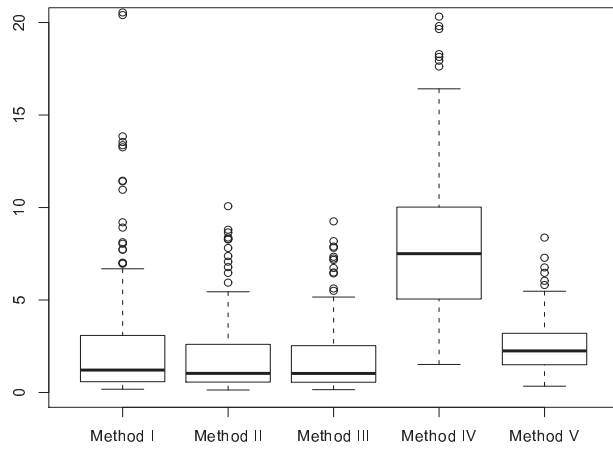


Figure 3.6: Simulation study. Case II. Box plot of the distribution of the $IMSE\beta$ for the estimated parameter functions on 200 repetitions given by Method I, II, III, IV, and V.

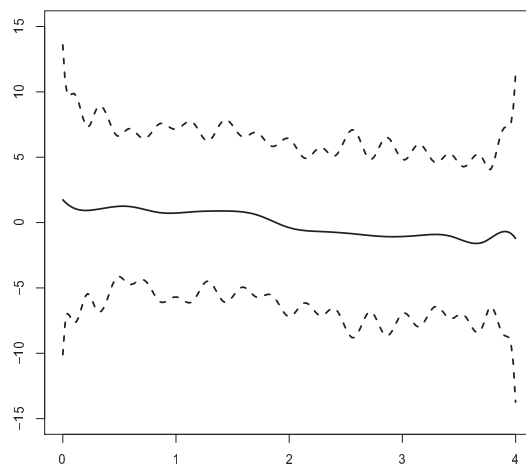


Figure 3.7: Simulation study. Case II. Mean of the functional parameters and the confidence bands (computed as the mean ± 2 the standard deviation) estimated by FLDA-PLS method over 200 simulations.

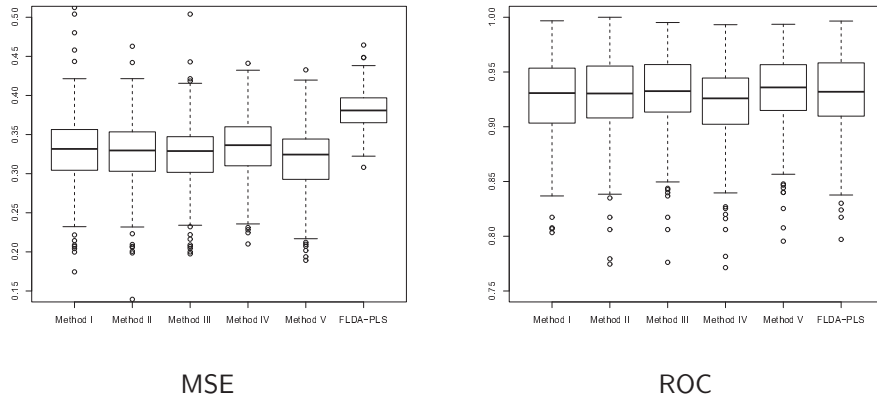


Figure 3.8: Simulation study. Case II. Area under ROC curve and MSE distribution (for the test samples of the 200 repetitions) given by Method I, II, III, IV, V, and FLDA-PLS.

formulation of functional PCA and the principal components are computed by maximizing a penalized sample variance that introduces a discrete penalty in the orthonormality constraint between the principal components weight functions. In Method IV the P-spline penalty is used in the maximum likelihood estimation of the functional parameter in terms of functional principal components. On the other hand, direct P-spline likelihood estimation in terms of B-spline functions is also considered (Method V).

Two simulation studies were performed to test the ability of the proposed P-spline smoothing approaches to provide an accurate and smooth estimation of the functional parameter and a good classification performance. Leave-one-out cross validation and generalized cross validation are adapted to select the different parameters (smoothing parameter and number of principal components or basis functions) associated with the considered approaches. In the case of the P-spline approximation of the sample curves from equally spaced observations, a relatively large number of equally spaced basis knots is a good choice for the definition of the B-spline basis. The results provided by the different smoothing approaches are compared with the estimations provided by non-penalized FPCLoR on least squares approximation of sample curves with B-spline basis (Method I) and by the partial least squares estimation approach for functional linear discriminant analysis (FLDA-PLS).

From the simulation study it can be concluded that the estimation of the

functional parameter given by the P-spline approaches is much smoother than the one given by the non-penalized FPCLoR although in some cases Method IV gives worse results. In fact, Methods I and IV provide non-smooth estimations affected by high variability. The most accurate and smoothest estimations of the parameter function are provided by Methods II and III, based on P-spline estimation of functional PCA with B-spline basis. On the other hand, the estimations given by Method V are less accurate and oversmoothed. In relation to the forecasting ability of the proposed methodologies, Methods II and III provide the least prediction errors followed by Method V that also gives accurate results. The classification performance of all methods is very good, with Methods II, III, and V being the most competitive. On the other hand, the FLDA-PLS approach gives very high classification rates similar to methods II, III and V but its forecasting errors are much higher.

In summary, it can be concluded that the penalized approaches represented by Methods II and III are preferred because they provide the most accurate estimation of the parameter function and have the best forecasting and classification performance, with Method II having lower computational cost.

An intuitive explanation of the fact that Methods II and III provide better estimation of the functional parameter could be that these approaches develop a penalized smoothing of the sample curves before estimating the regression model and select the smoothing parameter according to the mean squared error with respect to the observed sample curves. This way, the smoothing of the curves provides an smoothed estimation of the functional parameter. On the other hand, the results given by method V are not so good because the roughness of the functional parameter is directly penalized in the ML estimation but the smoothing parameter is selected by minimizing the prediction error without taking into account the smoothness of the sample curves. Finally, Method IV gives the worst estimations of the functional parameter because this approach does not penalize the roughness of any of the functions involved in the analysis and the penalty is only on the regression coefficients in terms of the non-penalized principal components.

Penalized spline approaches for functional PLS regression

4.1 Introduction

The functional linear model for a scalar response (FLM) was one of the first regression models extended to the case of functional data. Theoretical aspects related with this model were studied in Cardot et al. (1999). Depending of the functional nature of the predictor or response variable, other functional linear models have been subject of intensive study in the recent literature on FDA. The case where the predictor is a vector or scalar and the response is functional was studied by Chiou et al. (2004). Functional analysis of variance was introduced to model the mean of a functional response in terms of a categorical variable (Cuevas et al., 2002). On the other hand, functional linear models where both predictor and response variables are functional were studied by Yao et al. (2005b) and Ocaña et al. (2008). Principal component prediction models, that can be seen as a particular case of these linear models, were first introduced to forecast a continuous time stochastic process on a future interval from its recent past (Aguilera et al., 1997, 1999).

The aim of this chapter is to improve the estimation of the functional parameter associated with the functional linear model for a scalar response when the predictor curves are smooth functions observed with error. In order

to solve the problems of high dimension and multicollinearity related with the estimation of the FLM model, different approaches based on functional principal component regression (PCR), partial least squares regression (PLSR) and/or roughness penalty estimation were proposed in the related literature.

Several smoothing estimation approaches based on penalizing the least squares criterion in terms of a B-spline expansion of the functional parameter and smoothed principal component regression were considered by Cardot et al. (2003) to estimate the functional linear model with scalar response. On the other hand, PLS was extended to the case in which the predictor variable is functional and the response is scalar (Preda and Saporta, 2005b). New theory and explicit formulation of partial least squares for functional data was developed in Delaigle and Hall (2012b). Functional singular component analysis was introduced as an extension of multivariate partial least squares, where both predictors and responses are multivariate, to the functional case, where both predictors and responses are functional (Yang et al., 2011). In order to smooth the estimation of the FLM model, two different PCR and PLSR approaches for functional data based on B-spline basis expansion of the functional parameter and discrete roughness penalty estimation were proposed in Reiss and Ogden (2007). The main difference between these two approaches is in the way of introducing the penalty: in the likelihood estimation of the model or in the construction of the PLS or principal components. These penalized estimation approaches did not consider the functional form of the sample paths and are based on multivariate linear regression of the response in terms of the matrix of discrete-time observations of the sample curves. A penalized version of multivariate PLS was also applied for the estimation of additive functional models in terms of B-spline expansions of the variables (Krämer et al., 2008).

More recently, functional PCR and functional PLS with basis expansion of the sample curves and the parameter function were compared with their multivariate versions on an extensive simulation study (Aguilera et al., 2010b). From this study, the authors concluded that although discrete and functional models have similar prediction ability, the functional models provide a more accurate estimation of the functional parameter with FPLS giving the best estimation. In practice, an important problem associated with the PLS estimation of the functional linear model is the lack of smoothness of the estimated functional parameter that makes very difficult the estimation of the relationship between the response and the predictor variables. To solve this

problem, two new penalized approaches for functional PLS are introduced in this chapter. The first introduces the penalty in the definition of the norm of PLS component weight functions (this type of penalty was introduced by Silverman (1996) for the case of regularized FPCA). The second considers a penalized estimation of the covariance between the response and the PLS components (see Rice and Silverman (1991) that introduced a similar penalized estimation in FPCA). Continuous (based on the integrated squared d -order derivative) and discrete (based on d -order differences between adjacent coefficients) penalties are considered in terms of basis expansions of the sample curves. The performance of these penalized FPLS approaches is tested and compared with non-penalized FPLS on a simulation study where P-splines penalties are applied from least squares approximation of the sample curves with a B-spline basis. An application with chemometric functional data measuring the NIR spectra of gasoline data is also developed.

4.2 Functional PLS

Let Y be a scalar random variable (scalar response) and X be a second order stochastic process $\{X(t) : t \in T\}$ (functional predictor) whose sample paths belong to the space $L_2(T)$ of the square integrable functions. Without loss of generality, we assume that $E[Y] = 0$ and $E[X(t)] = 0, \forall t \in T$. With the aim of predicting Y from $X = \{X(t) : t \in T\}$, a functional linear model (FLM) is considered, so that

$$Y = \beta_0 + \int_T X(t) \beta(t) dt + \epsilon, \quad (4.1)$$

where $\beta(t)$ is the functional parameter, β_0 is a constant, and ϵ independent errors with zero mean. It is known that the use of the least squares criterion to estimate this model yields an ill posed problem because of the Wiener-Hopf equation which does not have a unique solution (Saporta, 1981).

In practice, an additional problem of the functional linear model is that we only have discrete observations x_{ik} of each sample path $x_i(t)$ at a finite set of knots $\{t_{ik} : k = 0, \dots, m_i\}$. In order to solve this problem basis expansions of $X(t)$ and $\beta(t)$ are usually considered.

Let us consider a basis $\{\phi_1(t), \dots, \phi_p(t)\}$ and assume that the functional predictor admits the basis expansion

$$X(t) = \sum_{j=1}^p \alpha_j \phi_j(t). \quad (4.2)$$

Let us also assume that the functional parameter admits a basis representation given by

$$\beta(t) = \sum_{k=1}^p \beta_k \phi_k(t).$$

Then, the functional model given in Equation (4.1) becomes a multiple linear model for the response variable in terms of a transformation of the functional predictor basis coefficients. Thus,

$$Y = \beta_0 + (\Psi\alpha)' \beta + \epsilon,$$

with $\alpha = (\alpha_1, \dots, \alpha_p)'$ and $\beta = (\beta_1, \dots, \beta_p)'$ being the vectors of the basis coefficients of X and β , respectively, and Ψ being the matrix of inner products between the basis functions, $\Psi_{p \times p} = (\psi_{jk}) = \int_T \phi_j(t) \phi_k(t) dt$.

In order to reduce the infinite dimension of the functional predictor and to solve the multicollinearity problem associated with the estimation of the FLM, a reduced number of functional principal components (Aguilera et al., 1999; Cardot et al., 1999, 2007; Yao et al., 2005b) or functional PLS components (Preda and Saporta, 2005b; Reiss and Ogden, 2007; Aguilera-Morillo et al., 2012) can be used as predictor variables to provide an accurate estimation of the functional parameter.

The functional PLS regression (FPLS) of a real random response Y in terms of a functional predictor $X = \{X(t) : t \in T\}$ is an iterative procedure, where the first PLS component $t_1 = \int_T X(t) w_1(t) dt$ is achieved by solving the following maximization problem

$$\begin{aligned} \max_w \quad & Cov^2 \left(\int_T X(t) w(t) dt, Y \right) \\ & \|w\|^2 = 1 \end{aligned} \quad (4.3)$$

Now, let define the operators

$$\begin{aligned} \mathcal{C}_{YX} : L^2(T) &\rightarrow \mathbb{R} \\ f &\rightarrow x = \int_T Cov(X(t), Y) f(t) dt \\ \mathcal{C}_{XY} : \mathbb{R} &\rightarrow L^2(T) \\ x &\rightarrow f(t) = Cov(X(t), Y) x, \forall t \in [0, T]. \end{aligned}$$

Denoting by $\mathcal{U}_X = \mathcal{C}_{XY} \circ \mathcal{C}_{YX} : L^2(T) \rightarrow L^2(T)$,

$$\begin{aligned}\mathcal{U}_X(w)(t) &= [\mathcal{C}_{XY}(\mathcal{C}_{YX}(w))](t) \\ &= [\mathcal{C}_{XY}(\text{Cov}(\langle X, w \rangle, Y))](t) \\ &= \text{Cov}(X(t), Y) \text{Cov}(\langle X, w \rangle, Y).\end{aligned}$$

Then,

$$\begin{aligned}\langle \mathcal{U}_X(w), w \rangle &= \langle \{\text{Cov}(X(\cdot), Y) \text{Cov}(\langle X, w \rangle, Y)\}, w \rangle \\ &= \text{Cov}(\langle X, w \rangle, Y) \langle \{\text{Cov}(X(\cdot), Y)\}, w \rangle \\ &= \text{Cov}(\langle X, w \rangle, Y) \int_T \text{Cov}(X(t), Y) w(t) dt \\ &= \text{Cov}^2(\langle X, w \rangle, Y) \\ &= \text{Cov}^2(\int_T X(t) w(t) dt, Y).\end{aligned}$$

The problem (4.3) can be written as

$$\max_w \frac{\langle \mathcal{U}_X w, w \rangle}{\langle w, w \rangle}. \quad (4.4)$$

Then, the weight function associated with the first PLS component corresponds to the largest eigenvalue of $\mathcal{C}_{XY} \circ \mathcal{C}_{YX}$. That is

$$\mathcal{C}_{XY} \circ \mathcal{C}_{YX}(w) = \lambda w.$$

Let $X_0(t) = X(t)$, $\forall t \in T$ and $Y_0 = Y$. Then, the first PLS step is completed by ordinary linear regression of $X_0(t)$ and Y_0 on t_1 , where $X_1(t)$ and Y_1 are the corresponding residuals so that

$$\begin{aligned}X_1(t) &= X_0(t) - p_1(t) t_1, \quad t \in T \\ Y_1 &= Y_0 - c_1 t_1.\end{aligned}$$

The second, and in general the h -th PLS component is given by

$$t_h = \int_T X_{h-1}(t) w_h(t) dt,$$

where $w_h(t)$ is obtained by solving the following problem

$$\begin{aligned}\max_w & \text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right) \\ & \|w\|^2 = 1\end{aligned}$$

The weight function associated with the h -th PLS component corresponds to the largest eigenvalue of $\mathcal{C}_{XY}^{h-1} \circ \mathcal{C}_{YX}^{h-1}$. That is

$$\mathcal{C}_{XY}^{h-1} \circ \mathcal{C}_{YX}^{h-1}(w) = \lambda w,$$

with \mathcal{C}_{XY}^{h-1} and \mathcal{C}_{YX}^{h-1} being the cross-covariance operators of $X_{h-1}(t)$ and Y_{h-1} . Finally, the h -th PLS step is concluded with the linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h and obtaining of the corresponding residuals

$$\begin{aligned} X_h(t) &= X_{h-1}(t) - p_h(t) t_h, \quad t \in T \\ Y_h &= Y_{h-1} - c_h t_h, \end{aligned}$$

where $p_h(t) = (\mathbb{E}(X_{h-1}(t) t_h) / \mathbb{E}(t_h^2))$ and $c_h = (\mathbb{E}(Y_{h-1} t_h) / \mathbb{E}(t_h^2))$.

4.2.1 Basis expansion estimation

Let us consider the basis expansion of $X(t)$ given by Equation (4.2) and the following basis expansion for the weight functions:

$$w(t) = \sum_{j=1}^p w_j \phi_j(t). \quad (4.5)$$

Then,

$$\begin{aligned} \langle \mathcal{U}_X(w), w \rangle &= Cov^2(\int_T X(t) w(t) dt, Y) \\ &= Cov^2(w' \Psi \alpha, Y) \\ &= w' \Psi \sigma \sigma' \Psi w \\ &= w' \tilde{\mathcal{U}} w, \end{aligned}$$

where $w = (w_1, \dots, w_p)'$ is the vector of basis coefficients of $w(t)$, $\tilde{\mathcal{U}} = (\Psi \sigma)(\Psi \sigma)'$, with $\sigma = (\sigma_1, \dots, \sigma_p)'$, so that $\sigma_j = \mathbb{E}(\alpha_j Y)$, and $\langle w, w \rangle = w' \Psi w$, with Ψ being the matrix of inner products between the basis functions.

Therefore, the maximization problem (4.4) can be written as follows

$$max_w \frac{w' (\Psi \sigma \sigma' \Psi') w}{w' \Psi w}. \quad (4.6)$$

Let us consider now the decomposition $\Psi = (\Psi^{1/2}) (\Psi^{1/2})'$. Then,

$$w' \Psi w = w' (\Psi^{1/2}) (\Psi^{1/2})' w = \tilde{w}' \tilde{w},$$

with $\tilde{w} = (\Psi^{1/2})' w$ ($w = (\Psi^{-1/2})' \tilde{w}$). This way, the maximization problem (4.6) turns into

$$\max_w \frac{\tilde{w}' (\Psi^{1/2})' \sigma \sigma' (\Psi^{1/2}) \tilde{w}}{\tilde{w}' \tilde{w}},$$

and the associated eigenvalue problem would be

$$(\Psi^{1/2})' \sigma \sigma' (\Psi^{1/2}) \tilde{w} = \lambda \tilde{w}. \quad (4.7)$$

The weight function associated with the first PLS component is given by $w_1 = (\Psi^{-1/2})' \tilde{w}_1$, with \tilde{w}_1 being the eigenvector associated with the largest eigenvalue of that problem. Then, the first PLS step is completed by ordinary linear regression of $X_0(t) = X(t)$ and $Y_0 = Y$ on t_1 , denoting by $X_1(t)$ and Y_1 the corresponding residuals.

The second, and in general the h -th PLS component $t_h = \int_T X_{h-1}(t) w_h(t)$ is obtained by the following problem

$$\max_w \frac{\tilde{w}' (\Psi^{1/2})' \sigma_{h-1} \sigma'_{h-1} (\Psi^{1/2}) \tilde{w}}{\tilde{w}' \tilde{w}},$$

where $\sigma_{h-1} = (\sigma_{h-1_1}, \dots, \sigma_{h-1_p})'$, with $\sigma_{h-1_j} = \mathbb{E}(\alpha_{h-1_j} Y_{h-1})$. The eigenvalue problem associated with this maximization problem is

$$(\Psi^{1/2})' \sigma_{h-1} \sigma'_{h-1} (\Psi^{1/2}) \tilde{w} = \lambda \tilde{w}. \quad (4.8)$$

Then, the weight function associated with the h -th PLS component is given by $w_h = (\Psi^{-1/2})' \tilde{w}_h$, with \tilde{w}_h being the eigenvector associated with the largest eigenvalue of that problem.

Finally, the h -th PLS step is concluded with the linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h , obtaining the corresponding residuals $X_h(t)$ and Y_h .

In general, by considering Equations (4.7) and (4.8), it can be concluded that FPLS is equivalent to an ordinary PLS of Y on the matrix $(\Psi^{1/2})' \alpha$ (Aguilera et al., 2010b).

Because of the relationship between the predictor and the response variable can be interpreted in terms of a functional parameter $\hat{\beta}(t)$, in practice, it is very important to obtain an accurate estimation of the parameters of a functional linear regression model. When we use a reduced number of functional PLS components as predictor variables of a FLM, the main problem comes from the estimated functional parameter. If we are working with noisy

data, the FPLS regression provides a non-smoothed functional parameter that could be difficult to interpret. In order to solve this problem, in this chapter the smoothness of $\hat{\beta}(t)$ is controlled by introducing a roughness penalty. Two different versions about penalized FPLS are shown below.

4.3 Penalized functional PLS

In this chapter, we investigate conditions under which applying smoothing in the new framework will allow an improvement in accuracy of the functional PLS components weight. Two different versions about penalized FPLS are developed. The first one makes use of the methodology exposed in Silverman (1996) for regularized FPCA that introduces the penalty in the norm of the PLS weights. The second one introduces the penalty in the covariance by following the penalized FPCA proposed in Rice and Silverman (1991). Before introducing both versions, we make a review about the penalty function which will be used in both penalized versions of FPLS.

4.3.1 Roughness penalty function

The approaches developed in this work will be based on a roughness penalty. In order to quantify the "roughness" of a function w on T , a roughness penalty such as $\int_T [D^d w(t)]^2 dt$ can be used. The first references about this kind of penalty can be seen in O'Sullivan (1986) who proposed to introduce a penalty in the second derivative of the curve $(\int_T w''(t) w''(t) dt)$. Thus, the flexibility of the fitted curve is restricted and the over-fitting is prevented.

By considering the basis function expansion of $w(t)$ given by Equation (4.5), the roughness penalty function is given by

$$\begin{aligned} PEN_d(w) &= \int_T [D^d w(t)]^2 dt = \int_T w' D^d \phi(t) D^d \phi'(t) w dt \\ &= w' [\int_T D^d \phi(t) D^d \phi'(t) dt] w \\ &= w' P_d w, \end{aligned} \quad (4.9)$$

where $w = (w_1, \dots, w_p)'$ is the vector of basis coefficients of $w(t)$ and P_d the matrix of the cross inner product of the d-order derivatives of basis functions.

In many applications the data are smooth functions observed with error. In this case least squares approximation can be used to estimate the basis

coefficients of a basis expansion of the unobserved smooth sample functions. In order to approximate smooth functions, the most accurate basis to be used is B-splines basis. B-splines are constructed from polynomial pieces joined smoothly at a set of knots. Once the knots are given, B-splines can be evaluated recursively for any degree of the polynomial by using a numerically stable algorithm (see De Boor, 2001).

In Eilers and Marx (1996) the approximation of O'Sullivan was generalized, such that it could be applied in any context where regression on B-splines was useful. They proposed to work with a relatively large number of knots and a penalty based on d-order differences between coefficients of adjacent B-splines. This kind of penalty was known as P-spline. In that chapter, the relationship between the two penalties was shown. In this context, $P_d = (\Delta^d)^T \Delta^d$, with Δ^d the matrix of d-order differences between the adjacent basis coefficients. From now, we used the term P_d for both continuous and discrete penalty matrix.

4.3.2 FPLS by penalizing the norm

In order to smooth the PLS functions associated with the non-penalized FPLS, the smoothing is incorporated in the definition of the norm with respect to an inner product which takes into account the roughness of the functions.

Let us define the inner product

$$\langle w, g \rangle_\lambda = \langle w, g \rangle + \lambda[w, g], \quad (4.10)$$

where $\langle w, g \rangle$ is the classical scalar product, λ is the smoothing parameter and $[w, g]$ is the roughness term

$$[w, g] = \int_T D^d w(t) D^d g(t) dt.$$

Let us observe that this is a generalization of the standard inner product and norms of Sobolev (see Adams, 1975).

By introducing the roughness penalty defined in Equation (4.9) in the functional PLS criterion, the first PLS component,

$$t_1 = \int_T X(t) w_1(t) dt,$$

is obtained by solving the following maximization problem:

$$\max_w \frac{Cov^2 \left(\int_0^T X(t) w(t) dt, Y \right)}{\langle w, w \rangle_\lambda}.$$

Equivalently, and in terms of $\mathcal{U}_X = \mathcal{C}_{XY} \circ \mathcal{C}_{YX}$,

$$\max_w \frac{\langle \mathcal{U}_X w, w \rangle}{\langle w, w \rangle + \lambda[w, w]}. \quad (4.11)$$

In general, for any of the two considered penalties (continuous penalty based on the integral of the squared second derivative or discrete penalty based on the differences between adjacent coefficients of the basis), the first PLS component, t_1 , is obtained by solving the problem

$$\max_w \frac{w' \tilde{\mathcal{U}} w}{\langle w, w \rangle + \lambda PEN_d(w)}, \quad (4.12)$$

with $PEN_d(w)$ being the penalty of order d defined in Equation (4.9).

Let us consider now the basis expansions of $X(t)$ and $w(t)$ given by Equations (4.2) and (4.5), respectively. Then, $\langle w, w \rangle_\lambda = w' \Psi w + \lambda w' P_d w$, with P_d being the penalty matrix, and $w = (w_1, \dots, w_p)'$ the vector of basis coefficients of $w(t)$. It is easy to convert the initial problem (4.12) into a new problem given by

$$\max_w \frac{w' \tilde{\mathcal{U}} w}{w' \Psi w + \lambda w' P_d w} = \max_w \frac{w' \tilde{\mathcal{U}} w}{w' (\Psi + \lambda P_d) w}. \quad (4.13)$$

By assuming the decomposition $LL' = \Psi + \lambda P_d$ ($L = (\Psi + \lambda P_d)^{1/2}$), the maximization problem (4.13) is equivalent to

$$\max_w \frac{w' \tilde{\mathcal{U}} w}{w' (LL') w}.$$

Defining $L'w = \tilde{w}$ ($w = (L^{-1})' \tilde{w}$) the problem is reduced to

$$\max_w \frac{\tilde{w}' (L^{-1})' \Psi \sigma \sigma' \Psi' (L^{-1})' \tilde{w}}{\tilde{w}' \tilde{w}}. \quad (4.14)$$

By analogy with the non-penalized FPLS, the associated eigenvalues problem will be

$$(L^{-1})' \Psi \sigma \sigma' \Psi' (L^{-1})' \tilde{w} = \lambda \tilde{w}, \quad (4.15)$$

so that the weight function associated with the first penalized PLS component is given by $w_1 = (L^{-1})' \tilde{w}_1$, with \tilde{w}_1 being the eigenvector associated with the largest eigenvalue of that problem.

The first PLS step is completed by ordinary linear regression of $X_0(t) = X(t)$ and $Y_0 = Y$ on t_1 , and denoting by $X_1(t)$ and Y_1 the corresponding residuals.

In general, the h -th PLS component ($h > 1$) is given by

$$t_h = \int_0^T X_{h-1}(t) w_h(t) dt,$$

where $w_h(t)$ is obtained by solving the following problem

$$\max_w \frac{\text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right)}{\langle w, w \rangle_\lambda},$$

whose associated eigenvalue problem would be

$$(L^{-1})' \Psi \sigma_{h-1} \sigma'_{h-1} \Psi' (L^{-1})' \tilde{w} = \lambda \tilde{w}. \quad (4.16)$$

The h -th PLS step is concluded with the linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h , and the corresponding residuals $X_h(t)$ and Y_h .

From Equations (4.15) and (4.16), it can be concluded that this version of penalized FPLS is equivalent to a ordinary PLS of Y in terms of the random vector $L^{-1} \Psi \alpha$.

4.3.3 FPLS by penalizing the covariance

In order to get accurate nonparametric estimation of the mean and the covariance structure, a roughness penalty estimation was proposed in Rice and Silverman (1991). Based on this penalty, an alternative smooth version of PLS regression in the functional data context is proposed in this section. The first penalized PLS component

$$t_1 = \int_T X(t) w_1(t) dt,$$

is achieved by estimating w_1 from the next maximization problem:

$$\max_w \frac{\text{Cov}^2 \left(\int_0^T X(t) w(t) dt, Y \right) - \lambda \text{PEN}_d(w)}{\langle w, w \rangle}, \quad (4.17)$$

with $PEN_d(w)$ the general penalty defined in Equation (4.9), and λ the smoothing parameter. Let us consider the basis expansion of $X(t)$ and $w(t)$ given by Equations (4.2) and (4.5), respectively. Then, the problem (4.17) can be rewritten as

$$\max_w \frac{w' \tilde{\mathcal{U}} w - \lambda w' P_d w}{w' \Psi w} = \max_w \frac{w' (\tilde{\mathcal{U}} - \lambda P_d) w}{w' \Psi w}. \quad (4.18)$$

Taking into account $\Psi = (\Psi^{1/2})(\Psi^{1/2})'$ and defining $\tilde{w} = (\Psi^{1/2})' w$ ($w = (\Psi^{-1/2})' \tilde{w}$), the problem (4.18) turns into this one

$$\max_w \frac{\tilde{w}' (\Psi^{-1/2}) (\Psi \alpha \alpha' \Psi' - \lambda P_d) (\Psi^{-1/2})' \tilde{w}}{\tilde{w}' \tilde{w}},$$

so that, it is reduced to

$$(\Psi^{-1/2}) (\Psi \alpha \alpha' \Psi' - \lambda P_d) (\Psi^{-1/2})' \tilde{w} = \lambda \tilde{w}. \quad (4.19)$$

Then, the weight function associated with the first penalized PLS component is given by $w_1 = (\Psi^{-1/2})' \tilde{w}_1$, with \tilde{w}_1 being the eigenvector associated with the largest eigenvalue of this problem.

The first PLS step is completed by ordinary linear regression of $X_0(t) = X(t)$ and $Y_0 = Y$ on t_1 , denoting by $X_1(t)$ and Y_1 the corresponding residuals.

In general the h -th PLS component ($h > 1$) is given by

$$t_h = \int_0^T X_{h-1}(t) w_h(t) dt,$$

where $w_h(t)$ is obtained by solving the eigenvalue problem (4.19) and considering $\Psi \sigma_{h-1} \sigma_{h-1}' \Psi'$ and $w_h = (\Psi^{-1/2})' \tilde{w}_h$. The PLS step is concluded with the linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h , with $X_h(t)$ and Y_h being the residuals.

4.3.4 Sample estimation

In this section, the estimation of the three considered models is developed. Let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of the functional variable $X(t)$ and $\{y_1, y_2, \dots, y_n\}$ be a random sample of Y associated with it. The functional linear model is then expressed as

$$y_i = \beta_0 + \int_T x_i(t) \beta(t) dt + \varepsilon_i,$$

where $\{\varepsilon_i : i = 1, \dots, n\}$ are independent and centered random errors.

The estimation procedure of the parameter function $\beta(t)$ using the basis expansion approach consists of

1. Computing the basis expansion approximation of $\{x_i(t), i = 1, \dots, n\}$. The vector of basis coefficients of the i -th sample path is estimated by the least squares criterion in terms of a B-spline basis, so that $\hat{\alpha}_i = (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i$, with $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$. This basis expansion approximation is known as regression spline.
2. Computing the PLS components. The matrix of PLS components T for each method is given by
 - Non-penalized FPLS: $T = A\Psi^{1/2}V$
 - FPLS by penalizing the norm: $T = A\Psi(L^{-1})'V$ ($LL' = \Psi + \lambda P_d$)
 - FPLS by penalizing the covariance: $T = A\Psi(\Psi^{-1/2})'V = A\Psi^{1/2}V$

with V being the matrix comprising the columns of the eigenvectors $\tilde{w}_1, \dots, \tilde{w}_p$ associated with the t_1, \dots, t_p PLS components.

3. The estimated functional linear model of Y in terms of the first h PLS components is given by

$$\hat{Y}^h = \mathbf{1}\gamma_0 + T^h \hat{\gamma}^h = \mathbf{1}\gamma_0 + A\Psi \hat{\beta}^h,$$

where T^h is the matrix whose columns are the first h PLS components, $\hat{\gamma}^h$ is the vector of the regression coefficients of Y on T^h , and $\hat{\beta}^h$ the vector of basis coefficients of the estimated parameter function

$$\hat{\beta}^h(t) = \sum_{j=1}^p \hat{\beta}_j^h \phi_j(t).$$

Then, the $\hat{\beta}^h$ vector of coefficients estimated by each method is given by

- Non-penalized FPLS: $\hat{\beta}^h = (\Psi^{-1/2})'V^h \hat{\gamma}^h$
- FPLS by penalizing the norm: $\hat{\beta}^h = (L^{-1})'V^h \hat{\gamma}^h$
- FPLS by penalizing the covariance: $\hat{\beta}^h = (\Psi^{-1/2})'V^h \hat{\gamma}^h$

where V^h is the matrix comprising the columns of the first h eigenvectors $\tilde{w}_1, \dots, \tilde{w}_h$ associated with the t_1, \dots, t_h PLS components of each considered method.

4.3.5 Model selection

In order to select the optimum number q of PLS components and the smoothing parameter λ , two different criteria have been considered:

- Criterion 1: choosing simultaneously the values of h and λ that minimize the GCV error

$$GCVE(h, \lambda) = \frac{n - \text{tr}(H_\lambda^h)}{n} \times \frac{MSE_\lambda^h}{n - \text{tr}(H_\lambda^h)},$$

where $H_\lambda^h = T^h((T^h)'T^h)^{-1}(T^h)'$ is the hat matrix, with T^h the matrix comprising the columns of the first h PLS components (estimated with the smoothing parameter λ), and $MSE_\lambda^h = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{i(h,\lambda)})^2$ is the mean squared error of the model estimated with h PLS components and using the smoothing parameter λ , being $\hat{y}_{i(h,\lambda)}$ the i -th prediction of that model.

- Criterion 2: for each number of PLS components, choosing the value of λ that minimizes the $GCVE(h, \lambda)$. Then, the number of PLS components h is selected by minimizing the integrated mean squared error of the parameter function

$$IMSE\beta(h, \lambda) = \left(\frac{1}{T} \int_T (\beta(t) - \hat{\beta}_\lambda^h(t))^2 dt \right)^{1/2},$$

where $\hat{\beta}_\lambda^h(t)$ is the parameter function estimated with h PLS components and the smoothing parameter λ . This criterion can be computed only in simulations where the parameter function is known.

For non-penalized FPLS both criteria are reduced to select only the number of PLS components.

4.4 Simulation study

The ability of the proposed penalized FPLS approaches to predict the response and to provide an accurate estimation of the functional parameter is tested and compared in this section with simulated data.

4.4.1 Description

The simulation study developed in this chapter is based on the spectroscopic data set of gasoline described by Kalivas (1997). The gasoline data set consists of the NIR spectra of 60 gasoline samples measured in 2-nm intervals from 900 nm to 1700 nm (400 discrete observations for each sample curve). The NIR spectra of these gasoline samples are shown in Figure 4.1 (left).

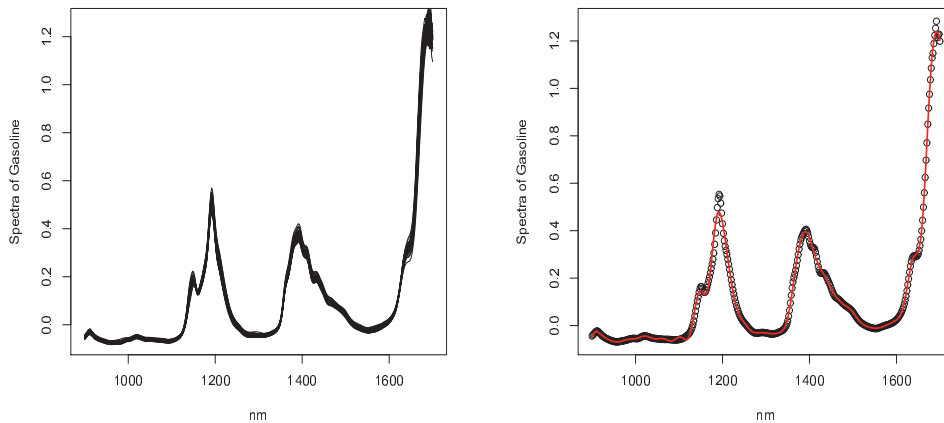


Figure 4.1: Simulation study. Spectrometric raw curves of 60 gasoline samples measured in 2-nm intervals from 900 nm to 1700 nm (left). Discrete observations (circles) and regression spline (red solid line) for one of the sample paths (right).

The parameter function $\beta(t)$ used for simulating the response variable Y of the functional linear model is a relatively smooth function

$$\beta(t) = 2 \sin(0.5\pi t) + 4 \sin(1.5\pi t) + 5 \sin(2.5\pi t), \quad t \in [0, 1],$$

used in Cardot et al. (2003) and Reiss and Ogden (2007) by transforming its

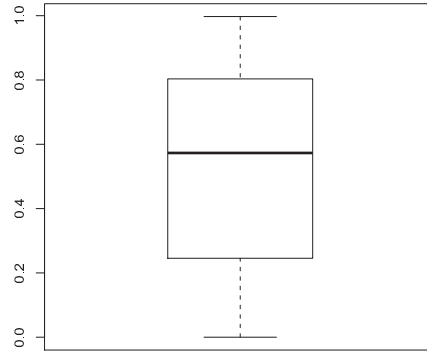


Figure 4.2: Simulation study. Box plot of the distribution of correlations between columns of matrix $A\Psi$.

domain to the spectra domain (see Figure 4.3 (left)).

After least squares approximation of the spectrometric curves and the functional parameter in terms of the cubic B-splines defined on 40 equally spaced knots in the interval $[900, 1700]$, the response values simulated in this work are given by

$$y_i = \int_{900}^{1700} x_i(t) \beta(t) dt + \varepsilon_i,$$

where ε_i ($i = 1, \dots, n$) are simulated independent random errors with normal distribution. The standard deviation of the errors, σ_ε , is chosen so that the squared multiple correlation coefficient of the true model equals 0.9 (Case I) and 0.7 (Case II). An example of regression spline (red solid line) for one of the sample paths (dotted line) is shown in Figure 4.1 (right).

The simulated response variable in matrix form would be given by $Y = A\Psi\beta + \varepsilon$, with A being the matrix of basis coefficients of the spectrometric curves, β the vector of basis coefficients of the parameter function and Ψ the matrix of inner products between the B-spline basis functions. It is well known that an important problem related with the estimation of this linear model is multicollinearity (high correlations between columns of its design matrix). This could produce inaccurate estimates of the functional parame-

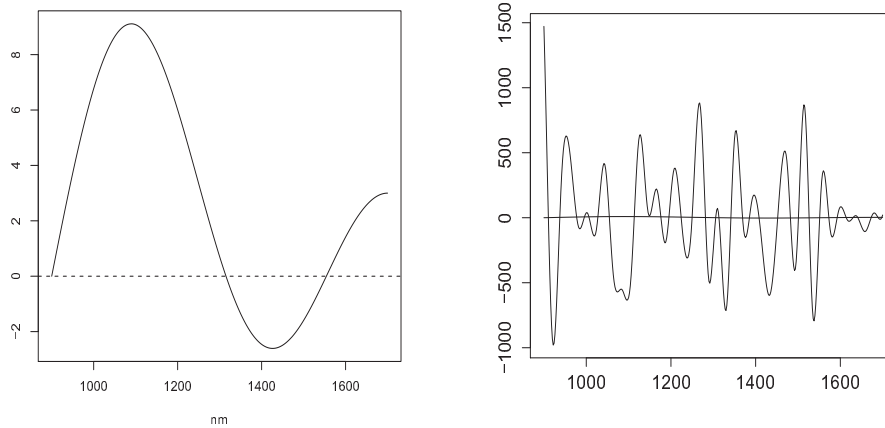


Figure 4.3: Simulation study. Parameter function used for simulating the response variable of the functional linear model (left) and the estimation provided by the functional logit model in terms of regression spline approximation of the spectrometric curves (right).

ter. The distribution of the correlations between the columns of the design matrix can be seen in Figure 4.2. As a consequence, the estimation of the functional parameter is really poor (see Figure 4.3 (right) for case I). This inaccurate estimation makes very difficult to interpret the relationship between the functional predictor and the response variable. This problem is solved in this chapter by applying different dimension reduction approaches based on taking an optimum set of functional PLS components as predictor variables.

The problem of lack of smoothness of the parameter function estimated by non-penalized FPLS (Method I) is solved by the two penalized estimations of FPLS regression introduced in this chapter: penalizing the norm and penalizing the covariance that will be called Method II and Method III, respectively, in this simulation study. The smoothing parameters associated with the two penalized FPLS versions are chosen by generalized cross validation (GCV). On the other hand, the optimal number of FPLS components in all compared methods were chosen by two different criteria: minimizing the GCV error (Criterion 1 denoted by CR1) and minimizing the integrated mean squared error with respect to the functional parameter ($IMSE_{\beta}$) (Criterion 2 denoted by CR2). In order to corroborate the good performance of the

penalized FPLS estimations proposed in this chapter, 100 repetitions of each simulation scheme (Case I and Case II) are carried out.

4.4.2 Discussion of results

As we have said before, the response values were simulated by fixing $R^2 = 0.9$ (Case I) and $R^2 = 0.7$ (Case II) for the simulation of the random errors associated with the functional linear model.

The means of the estimated parameter functions over the 100 simulations provided by Methods I (non-penalized FPLS), II (FPLS penalizing the norm), and III (FPLS penalizing the covariance) with criteria CR1 and CR2 used for model selection are displayed in Figure 4.4. Pointwise confidence bands computed as the sample mean ± 2 times the standard deviation at each time point are displayed in Figures 4.5 and 4.6 for Case I and II, respectively.

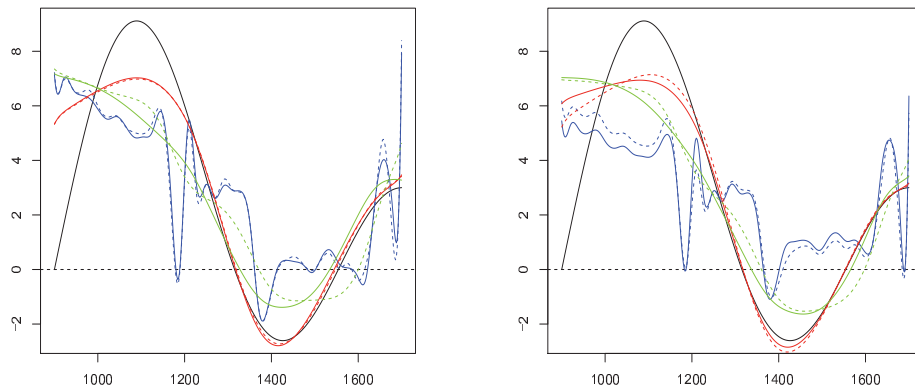


Figure 4.4: Simulation study. Case I (left) and Case II (right). Simulated parameter function (black), mean of the 100 parameter functions estimated by Method I (blue), II (red) and III (green), using Criterion 1 (solid line) and Criterion 2 (dashed line) for selecting the number of PLS components.

The box plots for the distribution of the integrated mean squared error with respect to the original parameter function ($IMSE_{\beta}$) are drawn in Figure 4.7 for the three considered methods and the two model selection criteria.

Let us observe that in Case II the variance of errors is higher and then the estimation of the parameter function is a bit less accurate and with higher variability than in Case I. This fact is reflected in the confidence bands for the mean functions that are wider for Case II.

According with the results related to the estimation of the parameter function, it can be seen that Method I (non-penalized FPLS) does not provide a fairly accurate estimation by showing a great lack of smoothness. Conversely, the penalized FPLS versions (Methods II and III) achieve a smooth parameter function estimation with the best being provided by Method II. Let us observe that Method III leads to over-smoothed parameter functions, and the widest confidence bands among the three compared methods. Anyway, Method I provides the worst estimation because loses control of smoothness with respect to the original parameter function, and then making very difficult their interpretation.

In order to test the prediction ability of the three considered methods, box plots for the distribution of the GCVE and MSPE (mean squared prediction error) on the 100 simulations are displayed in Figure 4.7.

Let us observe that the behavior of the distribution of both types of predictions errors is very similar. In the two cases, the penalized versions of FPLS (Methods II and III) provide slightly smaller prediction errors than the non-penalized FPLS approach (Method I). The prediction errors provided by Methods II and III are quite similar in Case I but in Case II the errors given by Method II are smaller.

To compare the degree of dimension reduction produced by the two model selection criteria, box plots for the distribution of the selected number of PLS components are drawn in Figure 4.7 for the two cases and the three FPLS approaches considered in this chapter. Let us observe that the penalized versions of FPLS regression (Methods II and III) require a somewhat smaller number of components than the non-penalized version (Method I). On the other hand, the number of PLS components selected by criteria CR1 and CR2 are similar except with Method III in which CR2 reduces the number of components providing a more accurate parameter function estimation than criterion CR1.

Let us take into account that in real applications the parameter function is unknown and CR2 criterion based on minimizing the $IMSE_{\beta}$ can not be used. It can be concluded that using GCV criterion for model selection is a good

option for predicting the response and estimating the parameter function except for Method III where the increment in the number of selected FPLS components worsen the functional parameter estimates slightly.

In order to see numerical differences among the results given by the FPLS regression Methods I, II and III, the model selection criteria CR1 and CR2, and the two different R^2 used in the simulation study, Table 4.1 summarizes the sample mean and the standard deviation of the errors $IMSE\beta$, GCVE and MSPE, and the number of PLS components for each of the eight possible combinations. The results in this table corroborate the previous ones given by Figures 4.4, 4.5, 4.6 and 4.7. Summarizing, it can be said that independently of the model selection criterion and the simulation scheme ($R^2 = 0.9$ or $R^2 = 0.7$), the more accurate estimation of the functional parameter is given by Method II.

With respect to the prediction errors, Methods II and III give similar results for $R^2 = 0.9$ with Method III providing slightly smaller error for $R^2 = 0.7$.

As expected, the significant differences between the non-penalized and penalized estimations of FPLS are not in their prediction ability but in their capacity to provide an accurate estimation of the functional parameter.

Table 4.1: Simulation study. Cases I and II. Sample mean and standard deviation related to the distribution of $IMSE_{\beta}$, GCVE, MSPE and number of PLS components for the optimum FPLS regression models estimated by Method I, II and III, with the number of predictors selected by CR1 and CR2 criteria on 100 simulations.

$R^2 = 0.9$							
		Criterion 1			Criterion 2		
	Method	I	II	III	I	II	III
IMSE $_{\beta}$	Mean	4.98	1.21	3.79	4.08	1.07	2.70
	sd	2.61	0.67	1.41	0.32	0.46	0.48
GCVE	Mean	0.18	0.16	0.14	0.18	0.16	0.16
	sd	0.04	0.03	0.03	0.04	0.04	0.03
MSPE	Mean	2.94	2.83	2.74	2.97	2.85	2.91
	sd	0.32	0.30	0.29	0.34	0.33	0.27
N $^{\circ}$ CPs	Mean	5.29	4.17	4.25	5.07	4.29	3.14
	sd	1.06	0.45	0.73	0.59	0.52	0.35

$R^2 = 0.7$							
		Criterion 1			Criterion 2		
	Method	I	II	III	I	II	III
IMSE $_{\beta}$	Mean	6.15	2.22	3.91	4.85	1.65	2.67
	sd	1.41	1.35	1.75	0.70	0.92	0.71
GCVE	Mean	0.64	0.60	0.51	0.68	0.62	0.52
	sd	0.12	0.11	0.09	0.12	0.11	0.09
MSPE	Mean	5.80	5.59	5.20	5.90	5.67	5.28
	sd	0.54	0.53	0.44	0.52	0.52	0.47
N $^{\circ}$ CPs	Mean	3.91	3.83	3.85	4.25	4.03	3.68
	sd	1.05	0.59	0.78	0.56	0.50	3.10

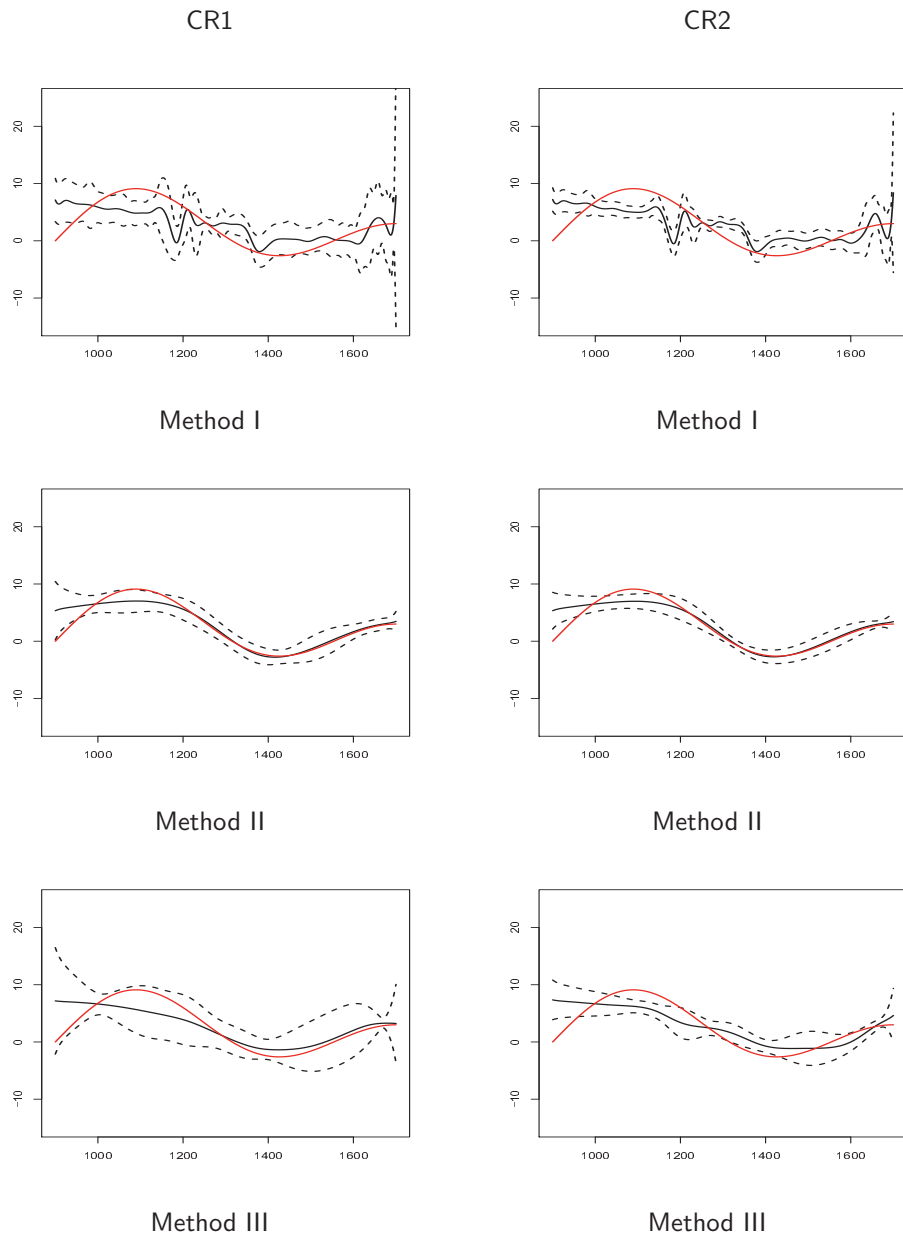


Figure 4.5: Simulation study. Case I. Simulated parameter function (red) and the mean of the 100 parameter functions estimated by Methods I, II, and III (black solid line) next to confidence bands (black dashed line) computed as ± 2 times the standard deviation at each time. The number of PLS components was selected by Criterion 1 (CR1, left column) and Criterion 2 (CR2, right column).

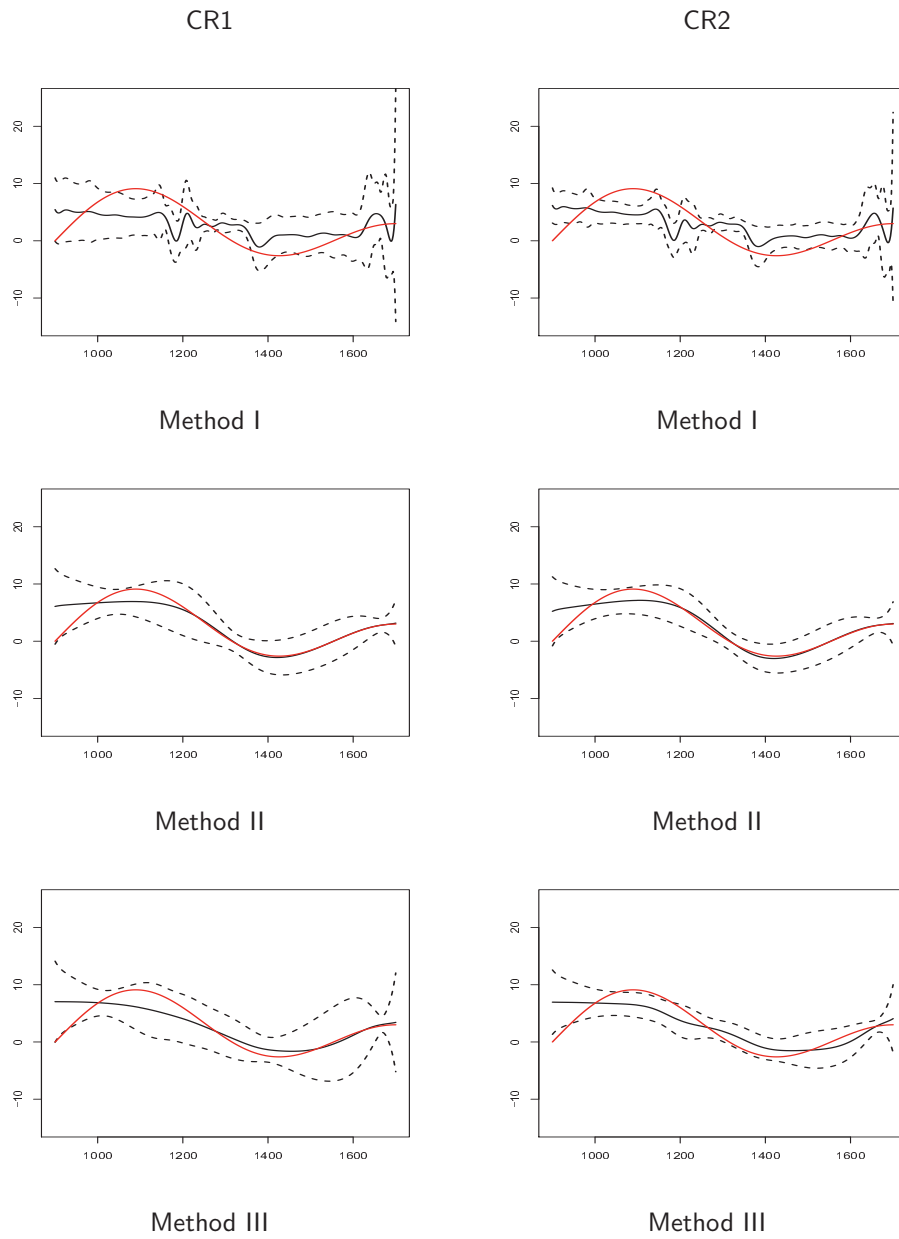


Figure 4.6: Simulation study. Case II. Simulated parameter function (red) and the mean of the 100 parameter functions estimated by Methods I, II, and III (black solid line) next to confidence bands (black dashed line) computed as ± 2 times the standard deviation at each time. The number of PLS components was selected by Criterion 1 (CR1, left column) and Criterion 2 (CR2, right column).

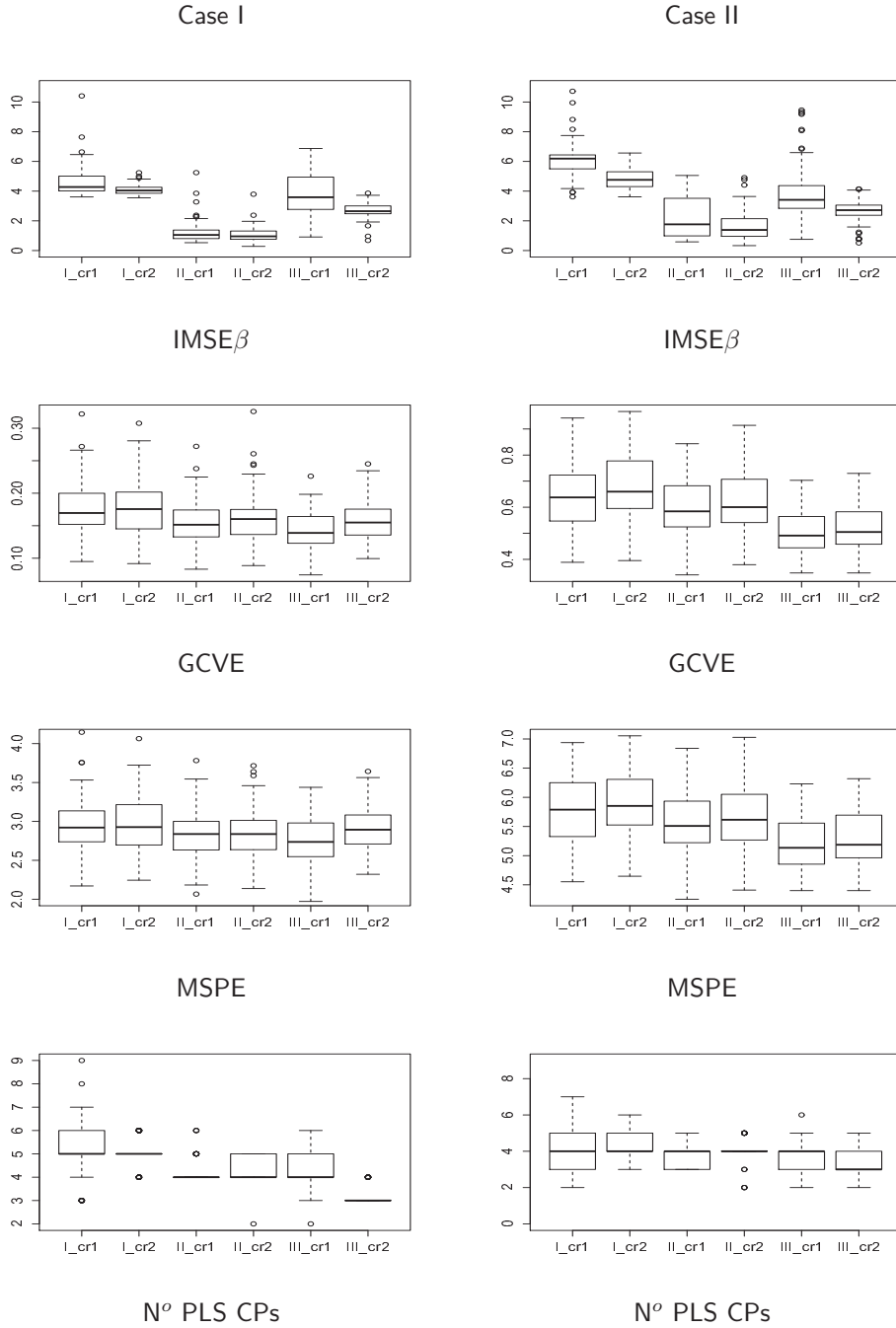


Figure 4.7: Simulation study. Case I (left) and Case II (right). Box plots related to the distribution of IMSE β , GCVE, MSPE number of PLS components for the FPLS regression models estimated by Method I, II and III, with the number of predictors selected by CR1 and CR2 on 100 simulations.

4.5 Real data application

In this section, the spectroscopic data set of gasoline described by Kalivas (1997) is considered again. Let $\{y_i : i = 1, \dots, 60\}$ be a sample of a scalar response variable related to the octane number of each gasoline sample. The aim is to forecast the octane number from the NIR spectra of 60 gasoline samples, measured in 2-nm intervals from 900 nm to 1700 nm. In order to test the results, the sample was divided into five sets of 12 gasoline samples, as in Reiss and Ogden (2007). Then, each of this subsets is taken as a test sample and the remaining 48 samples as a training sample to fit the model.

First, least squares cubic B-spline smoothing with 40 equally spaced knots of the spectra curves was carried out. After that, the three proposed methods (Method I, II and III) were applied. In this case, the parameter function is unknown, and then the number of PLS components for each method was chosen by GCV.

The mean of the parameter functions and the corresponding pointwise confidence bands estimated by Methods I, II and III for the five training sample are shown in Figure 4.8. The mean of the parameter functions estimated by the three proposed methods are overlaid in Figure 4.9. It is obvious that Method II provides the best estimations with less variability, following by Method III which gets an over smoothed parameter function with more variability than Method II. On the other hand, Method I provides a too noisy function.

In order to check the prediction ability of the different proposed methods, the mean squared prediction errors (MSPE) were computed on the five test samples. The sample mean and the standard deviation of the GCVEs, MSPEs, and the number of PLS components (N^o CPs) are summarized in Table 4.2.

Let us observe again that the penalized FPLS approaches provide smaller prediction errors than the non-penalized approach, with Method II giving the most accurate prediction. Respect to the dimension reduction, Method III requires the minimum number of predictors.

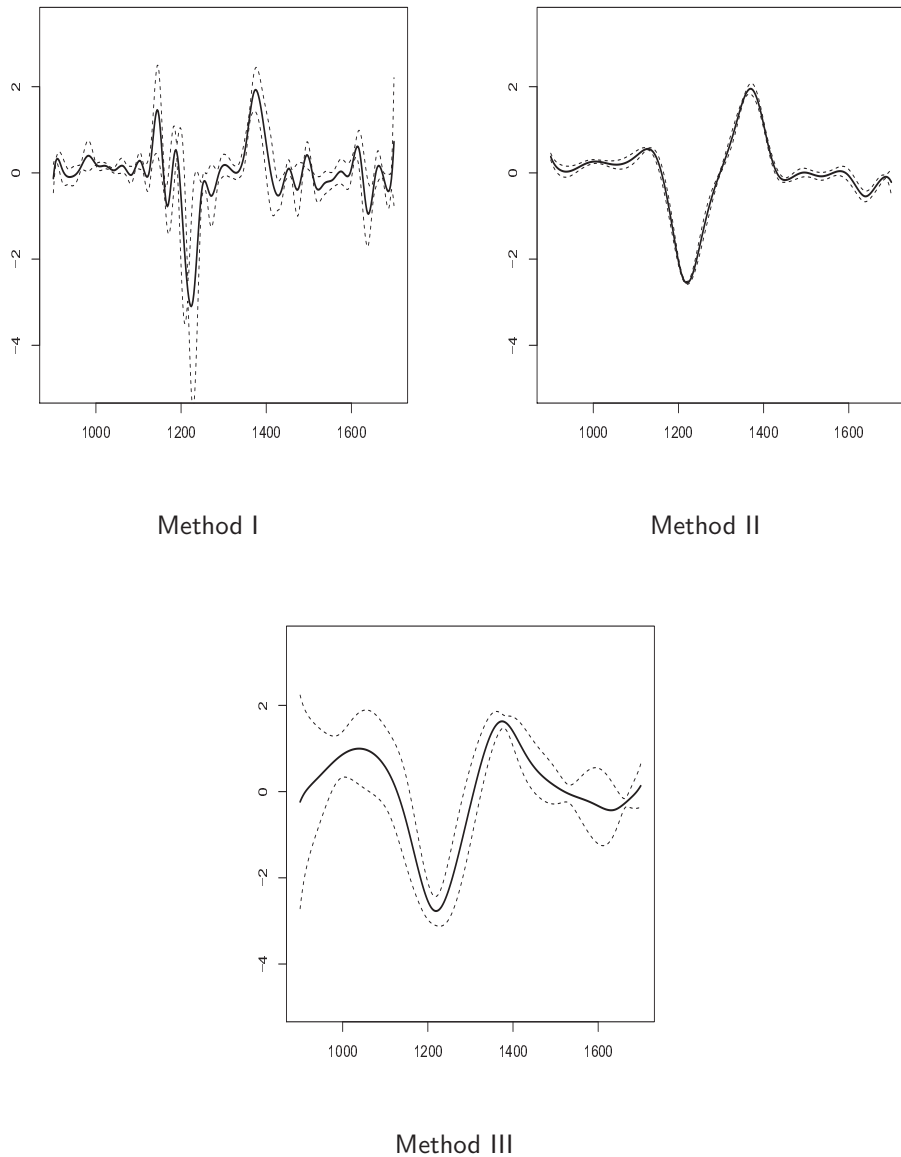


Figure 4.8: Application (gasoline data). Mean of the parameter functions estimated by Methods I, II, and III (solid line) for 5 training samples, next to the confidence bands (dashed line) computed as the mean ± 2 times the standard deviation at each time.

Table 4.2: Application (gasoline data). Mean and standard deviation of the distribution of the MSPE and the number of PLS components estimated by Methods I, II and III, for five models.

	MSPE		N° CPs	
	Mean	S.D	Mean	S.D
Method I	0.1231	0.0246	7	2.51
Method II	0.1031	0.0214	6	1.22
Method III	0.1061	0.0213	4	0.55

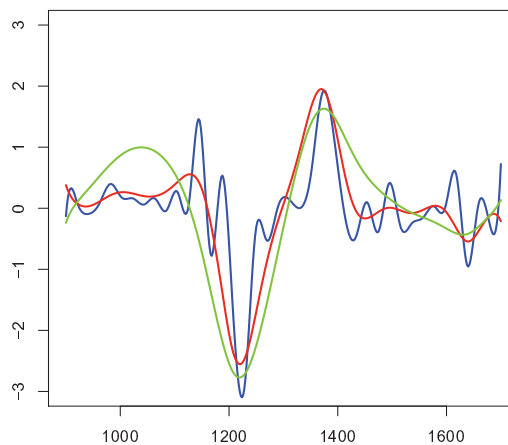


Figure 4.9: Application (gasoline data). Mean of the parameter functions estimated by Methods I (blue), II (red), and III (green), for 5 training samples.

4.6 Conclusions

The aim of this chapter is to improve the estimation of the functional parameter associated with the functional linear model for a scalar response when the predictor curves are smooth functions observed with error.

In order to solve the problem of high dimension and multicollinearity in the estimation of the functional linear model, and also to control the degree of smoothness of the estimated functional parameter, two different penalized approaches based on functional partial least squares regression (FPLSR) are

developed. The first approach introduces the penalty in the definition of the norm of the PLS component weight functions (Method II). The second one considers a penalized estimation of the covariance between the response and the PLS components (Method III). Discrete and continuous penalties can be used in terms of basis expansions of the sample curves.

Two different criteria based on minimizing the GCVE and the $IMSE_{\beta}$ (criterion 1 and 2, respectively) were adapted to select the different parameters (smoothing parameter and number of PLS components) associated with the considered approaches.

The performance of these penalized FPLS approaches was tested and compared with non-penalized FPLS by using least squares approximation of the sample curves with B-spline basis on a simulation study and an application with chemometric functional data measuring the NIR spectra of gasoline samples. In the simulation study two different schemes were considered so that $R^2 = 0.9$ and $R^2 = 0.7$.

From the simulation study, it can be concluded that the estimation of the functional parameter given by the penalized approaches is much smoother than the one given by the non-penalized FPLS. In fact, it can be said that independently of the model selection criterion and the simulation scheme ($R^2 = 0.9$ or $R^2 = 0.7$), the more accurate estimation of the functional parameter is given by Method II, because the estimations given by Method III are oversmoothed and present more variability. With respect to the forecasting performance, Methods II and III provide similar results, improving both the prediction ability of the non-penalized FPLS approach. The significant differences between the non-penalized and penalized estimations of FPLS are mainly in their capacity to provide an accurate estimation of the functional parameter. On the other hand, using GCV criterion for model selection is a good option for predicting the response and estimating the parameter function except for Method III where the increment in the number of selected FPLS components worsen the functional parameter estimates slightly.

In the application to the spectroscopic data set of gasoline, the aim was to forecast the octane number from the NIR spectra of 60 gasoline samples, and to get a good estimation of the functional parameter that explains the relationship between the response and the functional predictor. The results of this application corroborates that the penalized FPLS approaches have better forecasting performance and provide smoother estimated parameter than the

non-penalized approach, with Method II providing the best results. In both, simulation and application, Method III is which requires the minimum number of predictors.

Summarizing, Method II provides the best estimations of the functional parameter, achieving also a good forecasting performance, and using less predictors than the non-penalized FPLS approach.

P-spline estimation of functional classification methods for improving the quality in the food industry

5.1 Introduction

The aim of this chapter is to apply different functional classification methods to improve the quality in the production of the food industry. A major concern in the food industry is to offer good quality products that can be competitive in the market. This is the case of Danone biscuit manufacturer that intends to improve the quality of cookies by using only those flours that guarantee the best quality products.

There are several types of flour which are distinguished by their composition and way of extraction. The quality of a biscuit clearly depends on the type of flour used to make it. The main goal of this work is to classify cookies as good or bad in terms of the curves of resistance of the dough observed during the kneading process. This way, the flours that produce good cookies will be identified. A second purpose is to estimate the relationship between

the quality of cookies and the resistance of dough that allows to establish the main features of the curve of resistance for good cookies.

For each cookie, the resistance of dough is observed every two seconds. Let us observe that the data consist of repeated measurements of the resistance over time. Therefore, both longitudinal data analysis and functional data analysis (FDA) methodologies could be used to analyze them. It is known that data treated in FDA typically have high frequency (high number of equally spaced observations) whereas data treated in longitudinal data analysis are more typically sparse and irregularly sampled. Because of the high resolution of the data of resistance we will consider that the predictor data are functions and must be analyzed by using functional data analysis techniques (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horvath and Kokoszka, 2012). The perspectives and methods of functional data analysis and longitudinal data analysis for smoothing were contrasted and compared in Rice (2004).

There are many papers focus on the advantages of using FDA methodologies instead of the corresponding multivariate counterparts. The majority of them conclude that the multivariate analysis of the observed data fails in the case of unequally spaced sampling points by providing non accurate and unstable estimates (see Castro et al. (1986) for the case of principal component analysis and Aguilera et al. (1999) for principal component linear regression when both the predictor and the response data are functions). On the other hand, the estimation of a multivariate method is not always feasible from a set of observed longitudinal data. It is only recommended when data are observed at the same points for all individuals and the number of observations per individual is less than the sample size.

In order to solve the problem of classification of the curves of resistance in two groups (good or bad) we propose to apply and compared two different functional models. The first is the logit regression model (FLoM) (James, 2002) whose aim is not only to estimate a binary response variable from a functional predictor but also to provide a precise estimation of the relationship between the resistance of dough (functional predictor) and the quality of cookies (binary response). Interesting applications of this functional regression model were developed in different fields as medicine (Aguilera et al., 2008a) and environment (Escabias et al., 2005). The second is functional linear discriminant analysis (FLDA) (James and Hastie, 2001). Alternative nonparametric methods for curve classification were developed in Ferraty and Vieu (2003).

In order to solve the usual problems of high dimensionality and multicollinearity related with functional data analysis, dimension reduction techniques, such as functional principal component analysis (FPCA) and functional partial least squares (FPLS) regression are used to estimate the functional parameters in terms of basis expansions of the sample curves (Escabias et al., 2004, 2007; Preda et al., 2007). Functional principal component linear regression and FPLS were compared with their multivariate versions on an extensive simulation study with both equally spaced and irregularly spaced sampling points (Aguilera et al., 2010b). From this study, the authors concluded that the predictive ability of discrete and functional models is almost the same but the functional models provide a more accurate estimation of the functional parameter. A theorist study on near perfect classification of functional data that justify the very good practical performance of these dimension reduction methods was developed in Delaigle and Hall (2012a).

The curves of resistance of dough are smooth curves observed with error. Because of this, least squares approximation with B-spline basis is appropriate to approximate the true form of the curves. The problem is that regression splines do not control the degree of smoothness of the curves that depends on the position and number of the knots selected to construct the basis functions. As a result, the estimation of the functional parameters is not smooth enough. To solve the problem of lack of smoothness of the estimated functional parameter associated to the FLoM, four different P-spline-based approaches were introduced in Aguilera-Morillo et al. (2012). From the simulation results presented in this chapter, it was concluded that the estimation based on P-spline approximation of the sample curves (Eilers and Marx, 1996) is preferred because it provides the most accurate estimation of the parameter function and have the best classification performance with lower computational cost. Because of this, in this work, it is proposed to introduce this P-spline approximation in the FPLS estimation of FLDA. The classification results will be compared with the ones provided by the P-spline estimation of the functional principal component logit regression (FPCLoR) model and alternative FDA classifiers such as componentwise logit classification (Delaigle et al., 2012).

Inference results, such as confidence intervals for the functional parameters, based on the asymptotic normality of the likelihood estimators under the classical regularity conditions, are also provided for the FPCLoR model. Moreover, guidelines for the interpretation in terms of odds ratios and prin-

cipal components are also provided.

5.2 Smoothing the data

The aim is to classify a cookie as good or bad from the time evolution of the resistance (density) of the dough during the kneading process. This means that the main problem is to classify a set of curves, $\{x_i(t) : t \in T, i = 1, \dots, n\}$, (being i the sample curve index, and T the function support) representing the resistance of dough (functional data) according to a binary response Y that takes the value $Y = 0$ if the quality of the cookie is bad and $Y = 1$ if the quality is good.

In the data set we have a sample of 90 flours for which the resistance of dough was recorded every two seconds during the first 480 seconds of the kneading process. So, we have discrete-time observations of 90 sample curves of resistance of dough (50 for good and 40 for bad flours) that can be seen as independent realizations of a continuous-time stochastic process $X = \{X(t) : t \in [0, 480]\}$. The raw curves of resistance of dough are displayed in Figure 5.1 for good (left (a)) and bad (right (b)) curves.

The first step in functional data analysis is to reconstruct the true functional form of the sample curves from their discrete-time observations. The resistance of dough is a smooth curve measured with error. Least squares approximation on the basis of cubic B-spline functions was used in this chapter. This kind of approximation was used in previous works to solve the same problem with different type of functional data (Aguilera et al., 2008a, 2010b).

Let us remember that the sample functions related to the resistance of dough were observed at a finite set of time points $\{t_k = 2 \times k : k = 0, \dots, 240\}$. Then, the sample information is given by a set of vectors $\{x_i = (x_{i0}, x_{i1}, \dots, x_{i,240}) : i = 1, \dots, 90\}$, with x_{ik} being the value of the i -th sample path, $x_i(t)$, observed at the time t_k . That is, $x_{ik} = x_i(t_k) + \epsilon_{ik}$, with ϵ_{ik} being the smoothing error that follows a normal distribution.

Let us assume that the sample paths belong to a finite-dimension space generated by a basis $\{\phi_1(t), \dots, \phi_p(t)\}$ as in expression (1.2). Then, the vector of basis coefficients of each sample curve that minimizes the least squares error $(x_i - \Phi_i a_i)'(x_i - \Phi_i a_i)$, is given by $\hat{a}_i = (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i$, with $\Phi_i = (\phi_j(t_k))_{241 \times p}$ and $a_i = (a_{i1}, \dots, a_{ip})'$.

As we can see in Chapter 1, the problem of regression splines is that they

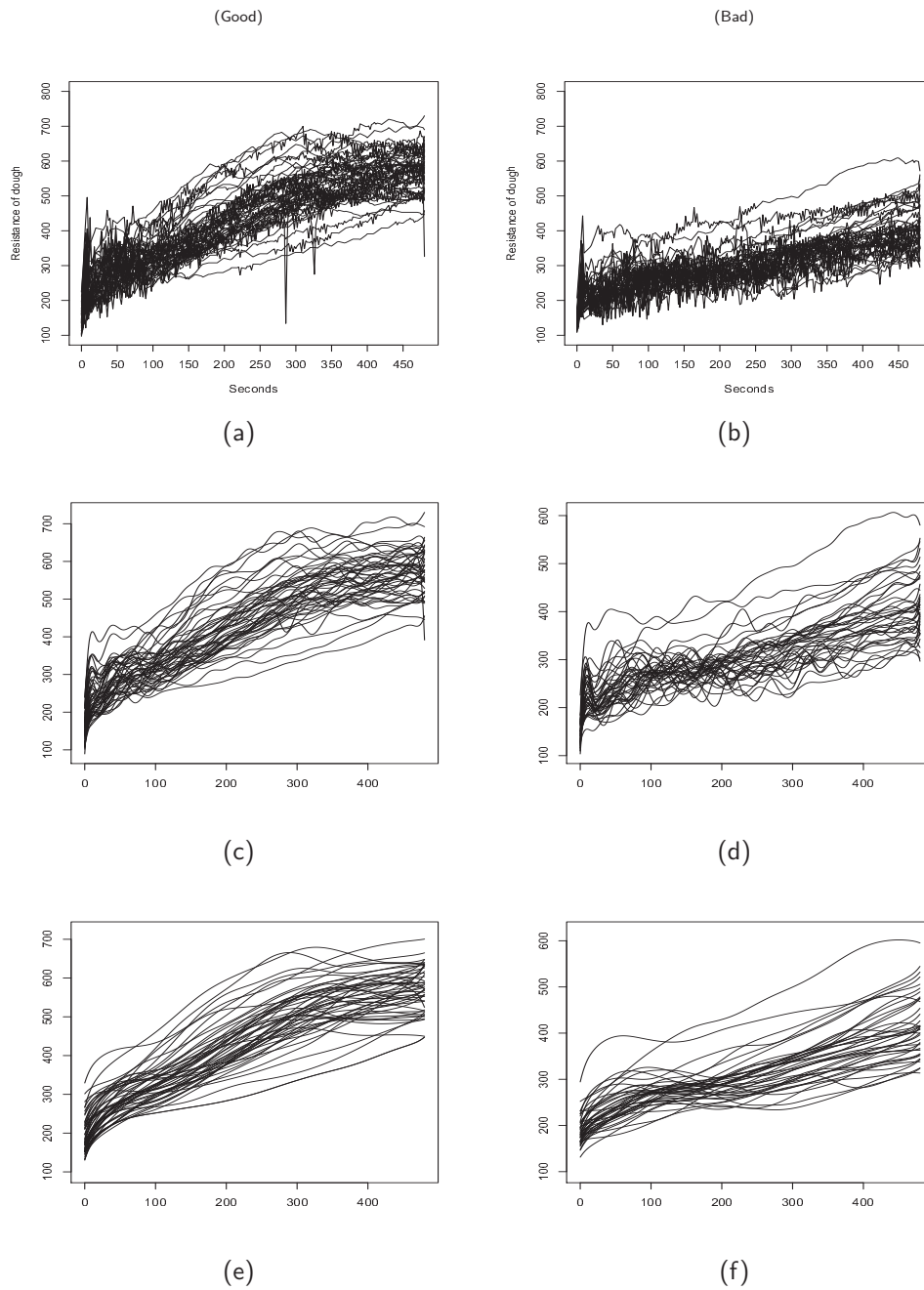


Figure 5.1: Curves of resistance of dough recorded at 240 seconds for 50 flours of good quality (left) and 40 flours of bad quality (right). Raw data ((a) and (b)), regression splines ((c) and (d)) and P-splines ((e) and (f)).

do not control the degree of smoothness of the estimated curve that depends on the number knots selected for defining the B-spline basis. This problem is solved in this chapter by using penalized splines that take into account the roughness of a curve by introducing into the least squares criterion a discrete penalty based on differences of order 2 between coefficients of adjacent B-splines (P-spline penalty).

The basis coefficients of the P-spline approximation of each sample curve computed by minimizing the penalized least squares error are given by expression (1.8).

The degree of smoothness of a P-spline is controlled by the smoothing parameter λ that measures the rate of exchange between fit to the data and variability of the function. In this application, the smoothing parameter is chosen by generalized cross validation. On the other hand, the choice and position of knots are not determinant when using P-splines and it is sufficient to choose a relatively large number of equally spaced basis knots (Ruppert, 2002; Currie and Durban, 2002). Taking into account the Ruppert's rule, 28 equally spaced knots were considered to define the cubic B-spline basis. Then, each curve $x_i = \{x_i(t) : t \in [0, 480]\}$ is represented by a set of 30 basis coefficients $a_i = (a_{i1}, \dots, a_{i30})$.

In Figure 5.1, the cubic regression splines fitted to good (c) and bad cookies (d) are displayed next to the cubic P-splines for good (e) and bad cookies (f). As an example, in Figure 5.2, the original sample data (dotted line), the cubic regression spline (dashed line) and the cubic P-spline (solid line) approximations with 28 equally spaced knots are displayed for a cookie of good (left) and bad (right) quality. It can be observed that the approximation provided by P-splines controls much better the smoothness of the curves.

5.3 Methodological aspects

The aim of this chapter is to predict a binary response variable Y (quality of cookies) from a functional predictor X (resistance of dough) that is equivalent to the problem of classification of the sample curves in the two groups defined by the response categories. To solve this problem we propose to apply two different functional models. The first is the FPCLoR model (Escabias et al., 2004) and the second is the Functional Linear Discriminant Analysis approach based on Functional Partial Least Squares (LDA-FPLS) (Preda et al.,

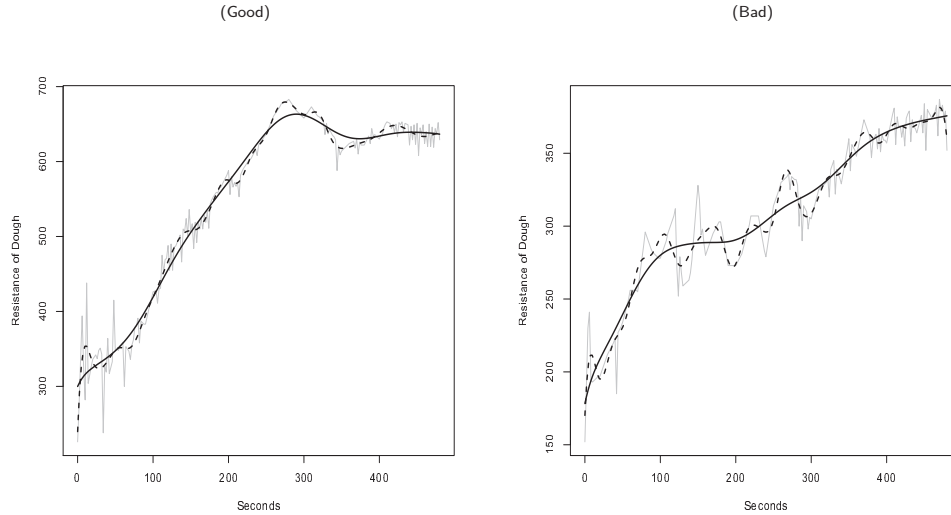


Figure 5.2: Original sample curve (dotted line), regression splines (dashed line) and P-spline approach (solid line) of a sample curve for good flour (left) and bad flour (right).

2007). Using functional PCA and PLS to estimate these models solves the usual problems of high dimensionality and multicollinearity related with functional data analysis. In both cases, a smooth version of these methodologies based on P-spline approximation of the sample curves is considered. These smoothed approaches provide a smoother functional parameter estimation easier to interpret. The theorist aspects related with these methodologies are summarized hereafter next to the basis ideas on componentwise classification (Delaigle et al., 2012) that will be applied for comparison purpose.

Let us consider the classification problem of a sample of functional observations $\{x_i(t) : t \in T; i = 1, \dots, n\}$ according to a related binary response $Y \in \{0, 1\}$ whose observations are denoted by $\{y_i : i = 1, \dots, n\}$. Let us also notice that the sample curves can be seen as observations of a second order stochastic process $X = \{X(t) : t \in T\}$ whose sample functions belong to the Hilbert space $L^2(T)$ of square integrable functions with the usual scalar product defined in Chapter 1 by Equation (1.1).

In what follows we will consider without loss of generality that both, the response and the predictor variables, are centered.

5.3.1 Functional principal component logit regression

Danone data set comprises a sample of 90 flours for which the resistance of dough was recorded every two seconds during the first 480 seconds of the kneading process. Then, a training and a test sample of 60 and 40 flours, respectively, were considered.

This data set presents several problems. On the one hand, the sample paths related to the resistance of dough of the cookies were observed at a finite set of 241 knots. On the other hand, it is impossible to estimate the infinite-dimensional functional parameter with a finite number of observations ($n = 60$). In order to solve the two problems at the same time, a functional estimation approach based on approximating the sample paths and the functional parameter in terms of basis functions is used.

Let us remember the FLoM is given by $y_i = \pi_i + \varepsilon_i$, $i = 1, \dots, n$, where $\{\varepsilon_i : i = 1, \dots, n\}$ are independent errors with zero mean, and $\pi_i = P[Y = 1 | \{x_i(t) : t \in T\}]$, and the associated logit transformations are expressed as

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

Let us consider the approximations in terms of p basis functions of the sample paths and the functional parameter given by expression (3.3). In this context, it is known that the FLoM turns into a multiple logit model whose design matrix is the product between the matrix of basis coefficients of the sample paths and the matrix of inner products between basis functions. Remembering the Chapter 3 of this thesis, the logit transformations in matrix form are given by Equation (3.4).

This model is affected by multicollinearity and high dimension of the functional predictor. The solution is to use a reduced set of q ($q < p$) functional principal components as predictor variables (Escabias et al., 2004).

In the Chapter 3 of this thesis, the FLoM in terms of functional principal components was introduced. By considering the basis expansion of both sample paths and parameter function, the functional principal components analysis (FPCA) is equivalent to the multivariate PCA of the product between the matrix of basis coefficients of the sample paths and the square root of the matrix of inner products between B-splines basis functions (Ocaña et al., 2007). Thus, the FLoM can be expressed in terms of the functional

principal components as in Section 3.3 of Chapter 3. The optimum number q of principal components will be selected by generalized cross validation and variability order.

Taking into account that the sample paths are smooth functions observed with noise, the approximations of the sample curves by using regression splines are not smooth enough. As a result, the functional parameter estimation provided by the FPCLoR model will be noisy and therefore difficult to interpret. For this reason, a smoothed estimation of the functional parameter in terms of P-spline approximation of the basis coefficients of the sample curves is proposed (see Subsection 3.3.2 for more details).

Functional parameter interpretation

In logit regression the exponential of the parameters are interpreted in terms of odds ratios. In this chapter we will consider the extension of this idea to the functional case proposed by Escabias et al. (2005).

Let l_i be the logit transformation for a specific functional observation $x_i(t)$ and l_i^* the logit transformation for this functional observation constantly increased in the period $[t_0, t_{0+h}]$. For these two logit transformations we have that

$$\exp(l_i^* - l_i) = \exp\left(K \int_{t_0}^{t_0+h} \beta(t) dt\right), \quad (5.1)$$

where K is a positive constant. This is an odds ratio, so that the odds of outcome $Y = 1$ is multiplied by this amount when the value of the functional observation is constantly increased in K units in a fixed interval $[t_0, t_{0+h}]$.

Inference on the functional parameter

The aim of this section is to obtain a confidence interval for the odds ratio given by Equation (5.1). Let $\hat{I} = \int_{t_0}^{t_0+h} \hat{\beta}(t) dt$ be the maximum likelihood estimator for $I = \int_{t_0}^{t_0+h} \beta(t) dt$, with $\hat{\beta}(t)$ being the maximum likelihood estimator of $\beta(t)$. In order to get a confidence interval for I , the variance of

\hat{I} is required by assuming its normal distribution. That is,

$$\begin{aligned}
 Var[\hat{I}] &= E \left[\hat{I} - E(\hat{I}) \right]^2 \\
 &= E \left[\int_{t_0}^{t_0+h} [\hat{\beta}(t) - E(\hat{\beta}(t))] dt \right]^2 \\
 &= \int_{t_0}^{t_0+h} \int_{t_0}^{t_0+h} E \left[[\hat{\beta}(t) - E(\hat{\beta}(t))] [\hat{\beta}(s) - E(\hat{\beta}(s))] \right] dt ds \\
 &= \int_{t_0}^{t_0+h} \int_{t_0}^{t_0+h} Cov(\hat{\beta}(t), \hat{\beta}(s)) dt ds.
 \end{aligned}$$

Let us now consider the basis representation of $\hat{\beta}(t)$ given by

$$\hat{\beta}(t) = \sum_{j=1}^p \hat{\beta}_j \phi_j(t) = \phi'(t) \hat{\beta},$$

with p being the number of basis functions, $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)'$ the vector of basis coefficients and $\phi(t) = (\phi_1(t), \dots, \phi_p(t))'$ the B-splines basis functions. Taking into account the vector of basis coefficients β estimated by smoothed FPCLo regression as $\hat{\beta} = F\hat{\gamma}$,

$$\begin{aligned}
 Cov(\hat{\beta}(t), \hat{\beta}(s)) &= \phi'(t) Cov(\hat{\beta}) \phi(s) \\
 &= \phi'(t) Cov(F\hat{\gamma}) \phi(s) \\
 &= \phi'(t) FCov(\hat{\gamma}) F' \phi(s),
 \end{aligned}$$

where $F = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times q}$, with G being the matrix whose columns are the eigenvectors of the covariance matrix of $A\Psi^{1/2}$. Then, the variance of the maximum likelihood estimator will be

$$Var[\hat{I}] = \int \int \phi'(t) FCov(\hat{\gamma}) F' \phi(s) dt ds,$$

where $Cov(\hat{\gamma})$ is the covariance matrix of the maximum likelihood estimator of the vector of parameters in terms of the principal components.

A $(1-\alpha)$ confidence interval for I is given by $\hat{I} \pm \hat{\sigma}(\hat{I}) z_{\alpha/2}$. Then, a $(1-\alpha)$ confidence interval for $\exp(I)$ can be computed by applying exponentials at both ends of the interval. In addition, a $(1-\alpha)$ pointwise confidence interval for $\beta(t)$ is given by $\hat{\beta}(t) \pm \hat{\sigma}(\hat{\beta}(t)) z_{\alpha/2}$, where $\hat{\sigma}(\hat{\beta}(t))$ is the standard deviation of $\hat{\beta}(t)$ given by the squared root of $Cov(\hat{\beta}(t), \hat{\beta}(t))$.

5.3.2 Functional linear discriminant analysis based on functional PLS regression

Linear discriminant analysis (LDA) in the functional data context aims to find linear combinations $\int_T X(t) \beta(t) dt$, $\beta \in L_2(T)$, such that the variance between classes is maximized with respect to the total variance

$$\max_{\beta} \frac{\text{Var}(E[X(t) | Y])}{\text{Var}(X(t))}.$$

Because of the equivalence between linear discriminant analysis and linear regression, the functional PLS regression approach (Preda and Saporta, 2005b,a) was used for classification purposes in Preda et al. (2007). The discriminant function is the coefficient function of the functional linear regression of Y on $\{X(t) : t \in T\}$, where Y is recoded as follows

$$\begin{aligned} Y &= -\sqrt{p_0/p_1} \quad \text{if } Y=1 \\ Y &= \sqrt{p_1/p_0} \quad \text{if } Y=0, \end{aligned} \quad (5.2)$$

with $p_0 = P[Y = 0]$ and $p_1 = P[Y = 1]$.

It is known that the estimation of the functional linear model

$$y_i = \beta_0 + \int_T x_i(t) \beta(t) dt + \varepsilon_i, \quad i = 1, \dots, n,$$

under least squares criterion is an ill-posed problem in the context of functional data (Cardot et al., 1999). One solution to this problem is to use dimension reduction approaches such as functional principal component regression and functional partial least squares (FPLS) regression. Both methodologies were compared on different simulated data sets concluding that their forecasting performance is similar but the estimated parameter function provided by FPLS regression is more accurate (Aguilera et al., 2010b). Because of this the FPLS approach will be used in this application.

As established in Chapter 4, the aim of FPLS regression is to regress Y on a set of uncorrelated random variables (FPLS components), in the linear space spanned by X , that take into account the correlation between the response Y and the functional predictor X . The procedure to compute the FPLS components was described in Chapter 4, Section 4.2.

By considering the basis expansion of both the sample paths and the discriminant function, FPLS is equivalent to multivariate PLS of the response

on the product of the matrix of basis coefficients of the sample curves and the squared root of the matrix of inner products between basis functions (Aguilera et al., 2010b).

As in the case of FPCLoR, the FPLSR model is reduced to multiple linear regression on a reduced set of PLS components, so that the vector of basis coefficients of the linear discriminant function is estimated as $\beta = F\gamma$, with G being the matrix whose columns are the PLS components loadings. The optimum number q of PLS components is also computed by generalized cross validation.

The estimated discriminant function is affected again by noise and then a smoothed estimation is proposed. As in FPCLoR, this estimation is based on the P-spline approximation of the sample curves with B-spline functions.

5.3.3 Componentwise classification

In order to extend the scale of comparison, a non-linear functional classifier introduced in Delaigle et al. (2012) is considered by comparing its performance with the proposed approaches.

Let $(x_1, y_1), \dots, (x_n, y_n)$ be independent and identically distributed data pairs corresponding to the sample information, where y_j is a class label taking values $\{0, 1\}$, and $\{x_i = (x_{i0}, x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$, with x_{ij} being the value of the i -th sample path, $x_i(t)$, observed at the time t_{ij} .

The componentwise classification approach consists of determining a relatively small number of points $\{t_1^*, \dots, t_p^* \in T\}$ that have important leverage for classification and applying a standard classification method on the vector $(X(t_1^*), \dots, X(t_p^*))$. A detailed study on the theoretical properties of the method and its behavior for different classifiers can be seen in Delaigle et al. (2012). In this chapter we will illustrate this approach for the classifier based on the logit regression model. The resulting logit-based componentwise classification approach has two main steps based on selecting the value of p and the position of the set of optimum time knots adaptively.

Choosing the optimal points in a given dimension

It is known that a full search taking into account, for successively higher values of r , all possible sequences $t_{(r)} = (t_1^*, \dots, t_r^*)$, can be feasible for $r = 1, 2$ or 3 , but becomes computationally costly for higher values of r . Because of this,

the algorithm is considered by combining a sequential and a refining part as follows:

1. Sequential part

Let us define for all possible set of r -vectors $t_{(r)} = (t_1^*, \dots, t_r^*)$ with $t_1^* < \dots < t_r^*$, the cross validation error rate for the logit regression model of Y on $(X(t_1^*), \dots, X(t_r^*))$ given by

$$\hat{e}(t_{(r)}) = \frac{1}{n} \sum_{i=1}^n C_i(t_{(r)}),$$

where $C_i(t_{(r)})$ is the logit classifier given by

$$C_i(t_{(r)}) = \begin{cases} 1 & \text{if } \hat{y}_i^{-i}(t_{(r)}) \neq y_i \\ 0 & \text{in other case,} \end{cases}$$

with $\hat{y}_i^{-i}(t_{(r)})$ being the estimation of y_i with the i -th data pair $(x_i(t), y_i)$ removed from the original sample.

- For $r = 1$, for all possible sets $t_{(1)} = t_j$, $j = 1 \dots, k$, the cross validation error rates $\hat{e}(t_{(1)})$ are computed and the one-dimensional point $t_{(1)}$ that provides the minimum error rate, \hat{t}_1 , is selected as the most important one-dimensional point for classification. Then, let us define $T_1 = \inf_{\{t_{(1)}\}} \hat{e}(t_{(1)}) = \hat{e}(\hat{t}_1)$.
- For $r \geq 2$, t_r^* is estimated as the value \hat{t}_r^* which, when adjoined to $\{\hat{t}_1^*, \dots, \hat{t}_{r-1}^*\}$, leads the smallest value of $T_r = \inf_{\{t_{(r)}\}} \hat{e}(t_{(r)}) = \hat{e}(\hat{t}_{(r)})$, and so on.

For the sequential part of the algorithm, it is recommended performing the search for each t_r^* on a grid of approximately 150 equally spaced points over the interval T , so that the space between any two selected points t_i and t_j , ($i \neq j$) will be at less 2 times the space between two adjacent points of the initial grid. If the sample paths are observed in less than 150 knots, the corresponding number of knots is considered instead of 150.

In the application developed in this work, the sample paths of resistance dough were recorded every two seconds during the first 480 seconds of the kneading process. The space between adjacent knots is 2 seconds

($\Delta t = 2$). Then, regarding to the point 1 of this algorithm, instead of 150 knots, 120 knots must be considered in order to get a space of $2\Delta t = 4$ seconds between adjacent knots.

2. Refining part

The sequential part, which was first used in a related functional problem by Ferraty et al. (2010), usually does not provide a consistent estimation of the optimal points. Because of the selected vector of optimum points is refined by constructing a neighbourhood around each point $\hat{t}_{(r)} = (\hat{t}_1^*, \dots, \hat{t}_r^*)$, and performing a full search over this point in that neighbourhood. Then, proceeding similarly in the following steps.

For $r \geq 1$, after the sequential part in which \hat{t}_r^* was adjoined to the points $\{\hat{t}_1^*, \dots, \hat{t}_{r-1}^*\}$, an optimum \hat{t}_r^* is chosen by taking shorter and shorter grids. That is, for each \hat{t}_j^* ($j = 1, \dots, r$) 20 neighbouring points equally spaced (by two times the space between adjacent knots of the initial grid) are considered when $r = 2, 3$ and 10 neighbouring points equally spaced for $r = 4$. For $r \geq 5$ only the sequential part is computed.

Choosing the optimal dimension p

The optimum number of points is selected as $p = \inf\{r : (1 - \rho)T_r \leq T_{r+1}\}$, where ρ denotes a predetermined small proportion. In our application we used $\rho = 0.1$ by following the recommendation in Delaigle et al. (2012).

5.4 Results

The good performance of the smoothed version of FPCLoR and LDA-FPLSR models is shown in this section. As mentioned in the introduction section, this work has two different aims. On the one hand, classifying the curves of resistance in two groups according to quality of cookies. On the other hand, obtaining an accurate estimation of the parameter functions that facilitates interpretation. From a classification point of view, the penalized versions of FPCLoR and LDA-FPLSR are compared with their non-penalized counterparts and alternative approaches such as componentwise classification on the logit model.

In order to test the classification ability of the considered approaches, the original sample of 90 flours was divided in a training sample of 60 flours and

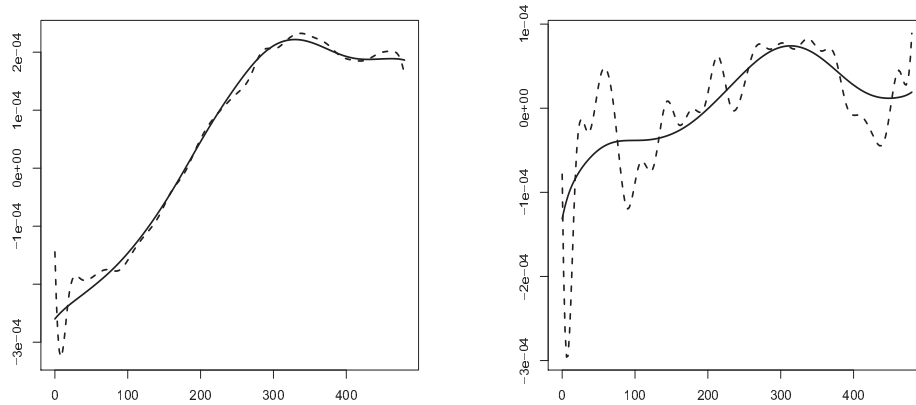


Figure 5.3: Means of the parameter functions estimated by FPCLoR (left) and the discriminant functions estimated by LDA-FPLSR (right), for unsmoothed (dashed lines) and smoothed versions (solid lines).

a test sample of 30 flours. To obtain general conclusions, one hundred re-sampling of the training and test samples were developed. In order to get the re-sampled data, i.i.d samples were randomly drawn from the original data, and then split into a training and a test set.

Firstly, the mean of the parameter functions estimated by FPCLoR (dashed line) and P-spline smoothed of FPCLoR (solid line) are displayed in Figure 5.3 (left). The mean of the discriminant functions estimated by LDA-FPLSR (dashed line) and P-spline smoothed LDA-FPLSR (solid line) are also shown in Figure 5.3 (right). In both cases, the penalized version provides the smoothest estimation of the corresponding function that makes the interpretation between the response and the predictor variables easier.

In regard to the prediction or classification ability of the considered methods, Table 5.1 presents the mean and the standard deviation (S.D.) of the areas under ROC curve (ROC) and the misclassification rates (MCR). It can be seen that the smoothed version of FPCLoR provides higher ROC and smaller MCR than the non smoothed FPCLoR. However, in the case of LDA-FPLSR, the values of ROC and MCR are very closed. In order to test if the observed differences between the MCR for the penalized and non-penalized

Table 5.1: Mean and sample standard deviation (S.D.) of the areas under ROC curve (ROC) and the misclassification rates (MCR) for 100 resamplings.

	ROC		MCR	
	Mean	S.D.	Mean	S.D.
FPCLoR	0.8792	0.2837	0.1626	0.2369
P-spline FPCLoR	0.9336	0.1884	0.1242	0.1861
LDA-FPLSR	0.9824	0.0146	0.0798	0.0406
P-spline LDA-FPLSR	0.9773	0.0171	0.0885	0.0915
CLoR	0.7141	0.1875	0.3333	0.1732

versions are statistically significant the one-tailed t-test for paired samples was performed. In the case of the FPCLoR model the p-value associated with the corresponding t-test was 0.01799. This means that the average MCR is lower for the P-spline smoothed FPCLoR than for the non-penalized version. In the case of LDA-FPLSR, the p-value is 0.6903. This means that the differences between the average of misclassification rates are not significant. This t-test was also applied for deciding if the differences observed between the average MCR provided by P-spline FPCLoR and LDA-FPLSR are significant. The p-value associated with this test was 0.0059 leading to the conclusion that the MCR is significantly lower for LDA-FPLSR than for FPCLoR. On the other hand, the componentwise logit classifier does not improve the classification ability of the proposed methods (FPCLoR and LDA-FPLSR) because its MCR is significantly higher. The most important advantage of the componentwise classifier is that greatly reduces the high dimensionality of the problem because rarely selected more than three or four time points. Let us observe from Figure 5.4 that the number of selected predictors variables (number of components selected by generalized cross validation in the case of FPCLoR and LDA-FPLSR, and number of selected time points in the case of componentwise logit regression (CLoR)) is much smaller for CLoR.

5.4.1 Interpreting the weight function

In this section, different interpretations of the parameter function associated to the FPCLoR model are carried out. The parameter functions estimated by FPCLoR (dashed line) and its P-spline smoothed version (solid line) are superposed in Figure 5.5 for one of the re-sampling. Let us observe that

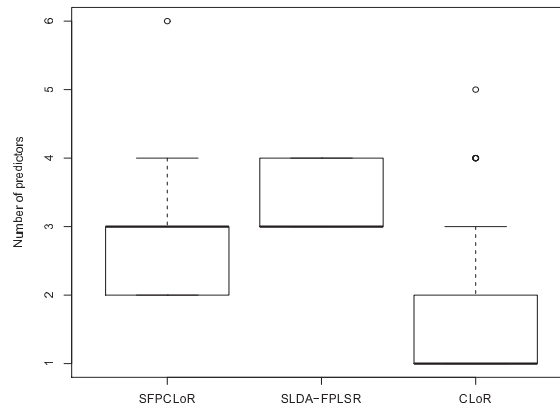


Figure 5.4: Box-plots for the distribution of the number of predictors selected for the three considered classifiers, smoothed FPCLoR (SFPCLoR), smoothed LDA-FPLSR (SLDA-FPLSR) and componentwise logit regression (CLoR), on one hundred re-sampling.

the weight function estimated by the P-spline FPCLoR approach is smoother than the non-penalized one and so more interpretable.

In Figure 5.6, the parameter function estimated by FPCLoR without smoothing (left) and with smoothing (right) are plotted next to 95% confidence bands computed in terms of the approximated pointwise confidence intervals introduced in Subsection 5.3.1 for each time point. It is obvious that the smoothed method presents less variability than the other one. Hereinafter all interpretations will be made on the smoothed function estimation of the parameter function.

In Figure 5.5, it can be seen that the point where the weight function is null is located in $t = 186$ seconds. In addition, the weights are negative in the early period of the kneading process (0, 186) and positive in the late period (186, 480). This means that flours with more resistance during the first 186 seconds of the kneading process have less probability of providing cookies with good quality meanwhile more resistance in the late period increase the probability of produce a good cookie. In other words, a good flour must have less resistance during the early period of the kneading process, and more resistance in the late period.

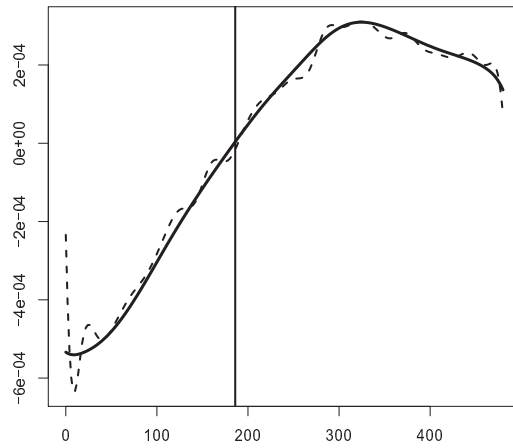


Figure 5.5: Parameter functions estimated by FPCLoR (dashed line) and P-spline smoothed FPCLoR (solid line).

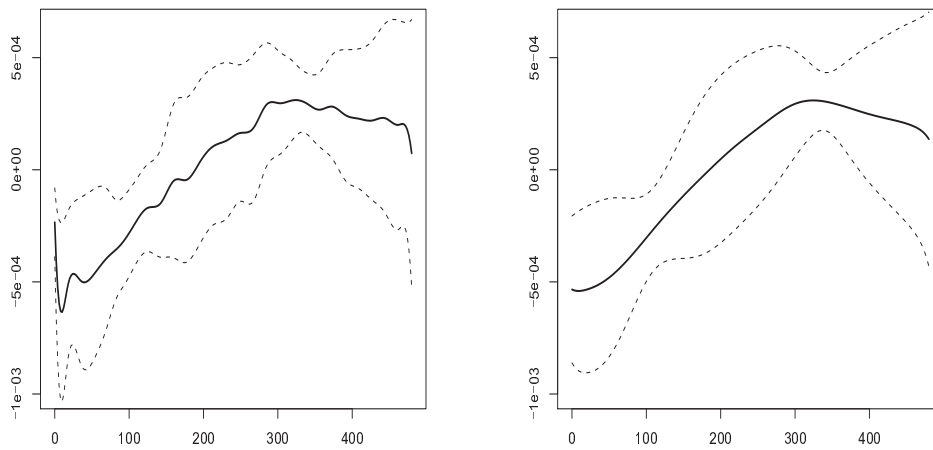


Figure 5.6: Parameter functions estimated by non smoothed (left) and smoothed FPCLoR (right) next to 95% pointwise confidence bands.

Odds ratio interpretation

A more precise interpretation of the odds ratios (introduced in Subsection 5.3.1) is based on a constant increase of the resistance of dough. The corresponding odds ratios for a one unit increase are given by

$$\widehat{OR}_1 = \exp\left(\int_0^{186} \hat{\beta}(t) dt\right) = 0.9433, \quad \widehat{OR}_2 = \exp\left(\int_{186}^{480} \hat{\beta}(t) dt\right) = 1.0663.$$

In both cases, the odd ratio is approximately 1. In order to check if it is significantly distinct to 1, 95% confidence intervals for the odds ratios were computed.

The confidence interval for OR_1 is given by (0.9104, 0.9774). Therefore, with a statistical significance of 5% the OR_1 takes value distinct to 1. In the same way, the confidence interval for OR_2 will be (1.0369, 1.0965). Therefore, with a statistical significance of 5% the OR_2 takes value distinct to 1.

Let us now consider a constant increase of 10 units in the resistance of the dough, so that

$$\widehat{OR}_1(\Delta X = 10) = \widehat{OR}_1^{10} = 0.56, \quad \widehat{OR}_2(\Delta X = 10) = \widehat{OR}_2^{10} = 1.9.$$

This means that if the resistance of dough is increased in 10 units (in a constant way) during the first 186 seconds of the kneading process, the odds of produce a good cookie is halved. On the other hand, if the same increase is made in a constant way during the period (186, 480), the odds of good cookie is doubled.

Principal component interpretation

After fitting the FPCLoR model according to the Equation (3.8), the first three principal components were selected by using generalized cross validation and variability order. Then, the fitted model is given by

$$\hat{l}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\xi}_{1i} + \hat{\gamma}_2 \hat{\xi}_{2i} + \hat{\gamma}_3 \hat{\xi}_{3i}, \quad (5.3)$$

with $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$ being the estimated coefficients and $\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3$ the first three principal components, which explains the 93.59%, 3.5% and 1.6% of the total variability, respectively. The results of the significance test on the estimated coefficients are shown in Table 5.2. Let us observe that only the parameter γ_1 is significantly different from zero.

Table 5.2: Significance test for the coefficients of the model.

	Estimate	Std. Error	z value	Pr(> z)
γ_0	1.2838	0.8936	1.44	0.1508
γ_1	0.0024	0.0007	3.35	0.0008
γ_2	0.0018	0.0019	0.99	0.3242
γ_3	0.0028	0.0028	0.97	0.3325

The first three principal component weight functions are displayed in Figure 5.7. The effect of adding and subtracting a suitable multiple of each factor loading to the sample mean function is shown in Figure 5.8. The dispersion graph between the first and the second principal component scores is displayed in Figure 5.9. It can be observed that the first principal component gives negative weights to the observations with response $Y = 0$ (bad cookies) and positive weights for $Y = 1$ (good cookies). On the other hand, the first principal component curve is always positive and represents the main features of the curve of resistance of a good cookie. This means that the main mode of variation of the resistance curves is associated with the quality of the cookies and allows us to identify the flours that produce good cookies.

Taking into account Equation (5.3), an increase of ξ_1 in one unit will cause that the odd in favor of the good quality of the cookies will be multiplied by $e^{\hat{\gamma}_1}$. Let us now consider the principal component decomposition

$$X(t) = \mu(t) + f_1(t) \xi_1 + f_2(t) \xi_2 + f_3(t) \xi_3,$$

with $f_j(t)$ being the j -th principal component curve. Then, $\Delta \xi_1 = 1$ produces an increase in $X(t)$ equal to $f_1(t)$. Therefore, if the resistance curves are increased according to f_1 , then the odd in favor of the good quality of the cookies will be multiplied by $e^{\hat{\gamma}_1} = 1.0024$. In addition, $e^{\hat{\gamma}_1}$ is significantly different from one with 5% statistical significance because the 95% confidence interval for $e^{\hat{\gamma}_1}$ given by (1.0012, 1.0035) does not contain the value 1. Finally, we can conclude that if we increase a resistance curve according to the first principal component curve, then the probability of producing a good cookie is increased.

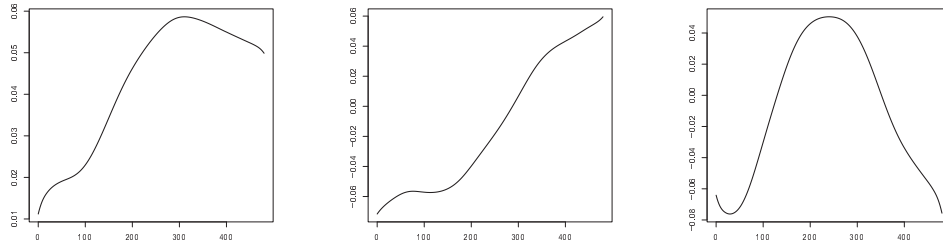


Figure 5.7: First eigenfunction (left), second eigenfunction (center) and third eigenfunction (right).

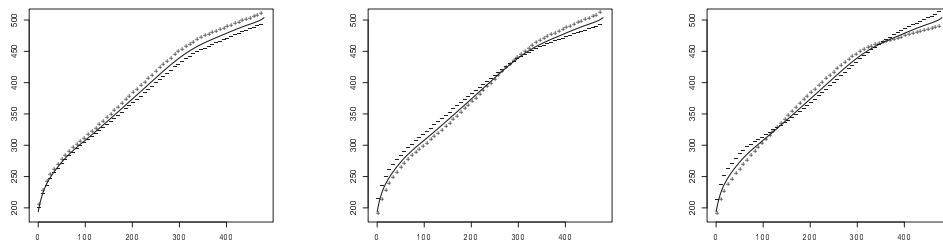


Figure 5.8: Mean resistance curves (solid line) and the effect of adding (+) and subtracting (-) a suitable multiple of each principal component curve.

5.5 Conclusions

Different functional classification approaches are applied and compared in this chapter to classify the quality of cookies (good or bad) in terms of the curves of resistance of dough during the kneading process. The aim of this application is to identify those flours that provide good cookies and improve the quality of the manufacturer's products by using only those flours that guarantee the best quality.

Two different classification methods, such as functional logit regression and functional linear discriminant analysis, are considered to classify a set of curves in the two groups defined by a binary response. A third method based on componentwise logit classification is also applied for comparison purposes.

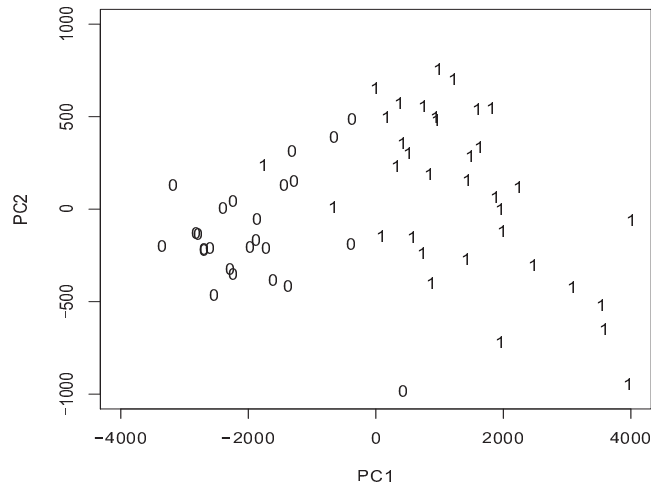


Figure 5.9: Dispersion graph between the first and the second principal component scores.

The first methodology allows not only to solve the classification problem but also to estimate the relationship between the response variable (quality of cookies) and the predictor variable (resistance of dough during the kneading process). The estimation of the proposed functional classification approaches is affected by several problems such as high dimension and multicollinearity. Estimation based on functional PCA and functional PLS is considered to solve these problems. Smoothed versions of these methodologies based on P-spline approximation of the sample curves with B-spline basis are introduced in this chapter to solve the problem of lack of smoothness of the estimated functional parameters. Inference on the estimated functional parameters and associated odds ratios is also carried out based on the asymptotic normality of the likelihood estimators.

From the statistical analysis of the results it can be concluded that the proposed functional methodologies (FPCLoR and LDA-FPLSR) have a high classification ability with LDA-FPLSR being the one that gives the highest area under ROC curve and the minimum misclassification rate. The smoothed versions of both approaches give more accurate and smooth estimations of the functional parameters which facilitates their interpretation. On the other

hand, the componentwise logit classifier does not improve the classification ability of the proposed methods because provides the smallest area under ROC curve and the highest misclassification error. The main advantage of this non linear classifier is that provides the highest reduction of dimension because rarely selected more than three or four time points that convey particular information for classification purpose.

Several interpretations of the functional parameter based on odds ratios and principal components are also proposed. To summarize, it can be concluded that good cookies have greater resistance of the dough in the late period and less resistance in the early period. The main features of the curve of resistance of good cookies were also identified by interpreting the first principal component curve.

Software and computational considerations

In this section the main computational considerations of this thesis are described. Note that the computational cost of this thesis is really high. Because of this, the largest computer simulations were run on a cluster of 30 blade servers each one with two Intel XEON *E5420* processors running at 2.5 GHz and with 16 GB of RAM memory. Each processor has four cores, and the experiments are carried out on virtualized Windows XP machines, each one with one virtualized processor and 1 GB of RAM memory.

All results presented in this thesis were performed by using the free software R for statistical computing and graphics (<http://www.r-project.org>). Specifically, two different versions of R software were used (2.10.1 and 2.15.2 version). This software is used by many researchers in FDA and there are some libraries related with the implementation of standard FDA methodologies. Moreover, the development of custom code for the developed algorithms not directly implemented in its libraries is allowed.

Then, a summary of the main libraries and R functions used for the development of this thesis is shown.

A.1 Main libraries

In this section, the main R libraries used to compute the methodology developed in this thesis are summarized.

- *fda* library
This library implements the techniques of functional data analysis described in Ramsay and Silverman (2005).
- *fda.usc* library
This library integrates and complements the *fda* package by carrying out exploratory and descriptive analysis of functional data, functional regression models with a scalar response, supervised and unsupervised classification methods and functional analysis of variance.
- *stats* library
This package contains functions for statistical methods such as principal components analysis and linear models, both used in this thesis.
- *design* library
Among other things, this library allows to fit a binary or ordinal logistic model by using ordinary or penalized maximum likelihood estimation.
- *mgcv* library
This library has been considered in order to make matrix calculations, such as the squared root of a matrix, by using either pivoted Choleski decomposition or singular value decomposition.
- *pls* library
This library was used to compute PLS regression.

A.2 Main functions

In this section, the main functions required to compute the methodology developed in this thesis are summarized. Graphics related with the mean of the parameter functions and pointwise confidence bands were computed by using the functions *func.mean* and *func.var* of the library *fda.usc*.

A.2.1 Smoothing with B-spline bases

The first step in FDA is to reconstruct the functional form of the sample curves from their discrete observations. The most usual way to solve this problem consists of assuming an expansion of each sample curve in terms of a basis of functions. In order to get the main basis functions, the following functions of the *fda* package were considered:

- Cubic B-splines basis: *create.bspline.basis(...)*.
- Fourier basis: *create.fourier.basis(...)*.
- Constant basis: *create.constant.basis(...)*.

In this thesis only cubic B-splines basis functions were used. Depending of the considered spline smoother (see Chapter 1 for more details), the basis coefficients are computed as follows:

1. Regression splines

The functional form is achieved with the function *data2fd(...)* of the *fda* library.

The curves are centered by using the function *center.fd(...)* of the *fda* package on the object obtained from the function *data2fd(...)*.

2. Smoothing splines

The continuous penalty matrix of order 2 used to compute the smoothing spline approximation of the sample curves is based on the integrated squared 2-order derivative of B-spline functions. It is computed by means of the function *bsplinepen(basis, Lfdobj=2)* of the *fda* library. The functional form are given by the function *smooth.spline*.

3. P-splines

The discrete penalty matrix of order 2 used to obtain the P-spline approximation of the sample curves is computed with the R code *diff(diff(diag(nnodos+2)))*, where *nnodos* is the number of basis knots.

A.2.2 Functional PCA

In order to carry out the different versions of penalized and non-penalized FPCA, the following libraries are required: *fda*, *mgcv*, *stats*.

Once the functional form of the sample paths is obtained, all versions of FPCA developed in this paper are reduced to a multivariate PCA of a design matrix which is achieved by mean of the function *princomp* of the library *stats*.

The design matrix of the different versions of FPCA are computed as follows:

- Non-penalized FPCA: the design matrix given by the product between the matrix of basis coefficients of the curves approximated by regression splines and the squared root of the matrix of inner products between basis functions.

The inner products between basis functions is computed with the function *inprod(...)* of the *fda* library. The squared root of that matrix is computed by using the function *mroot(...)* of the library *mgcv*.

- P-spline smoothed FPCA: the design matrix is given by the product between the matrix of basis coefficients of the curves approximated by regression splines, the matrix of inner products between basis functions and the inverse of a lower triangular matrix. The lower triangular matrix is obtained by a Choleski factorization of the sum of the matrix of the inner products between basis functions plus the smoothing parameter multiply by the P-spline penalty matrix. The Choleski factorization is computed with the function *chol(...)* of the package *base*.
- Functional PCA of P-splines: the design matrix is given by the product of the matrix of basis coefficients of the curves approximated by P-splines and the squared root of the matrix of inner products between basis functions.
- Regularized FPCA

The main difference between the computational algorithm for computing regularized FPCA and P-spline smoothed FPCA is the continuous penalty matrix that is based on the integrated squared 2-order derivative of B-spline functions.

A.2.3 Penalized functional PC logit regression

All the FPCLoR models developed in this thesis were reduced to ordinary logit regression of the binary response variable on the matrix comprising the columns of the first q functional principal component scores provided by different versions of penalized and non-penalized FPCA.

Once the principal components are computed, the logit model is achieved by means of the function `glm(formula, family=binomial(link="logit"),...)` of the R package *base*.

The optimal number q of functional principal component scores and the smoothing parameter (for penalized FPCLoR) was selected by the GCV algorithms described in Section by developing our own R code.

When FLoR with P-spline penalty in the maximum likelihood estimation is considered, the penalized estimation was provided by the function `lrm.fit(...)` of the library *design* after computing the P-spline penalty matrix.

A.2.4 Penalized functional PLS regression

Both, non-penalized FPLS and the first version of penalized FPLS (described in Section 4.3.2) are reduced to ordinary PLS regression of the response variable on an appropriate design matrix. Then, the PLS components are estimated by the function `pls(...)` of the *pls* library.

The vector of model parameters provided by the *pls* function is obtained in both cases with the `coef(pls(...), ...)` function. Then, the vector of basis coefficients of the functional parameter is obtained by transforming that vector as corresponds in each case.

The second version of penalized FPLS is not reduced to a multivariate PLS. Then, the penalized FPLS components were obtained by the iterative algorithm, where the required eigenvalues problems were computed with the function `eigen(...)` of the R package *base*. The linear regression models of the response variable on the penalized PLS components were computed with the function `lm(...)` of the library *stats*.

Conclusions and further research

Let us remember that the general objective of this thesis is to improve the estimation of FDA methodologies in the case of smooth functional data observed with error. In order to solve this problem, different penalized estimation approaches with B-spline basis expansions are proposed for functional PCA, functional principal component logit regression and functional PLS.

The performance of all the proposed penalized methodologies is studied on simulated and real data and the results compared with the corresponding non-penalized counterpart estimated in terms of B-spline basis expansions of the sample curves and the functional parameters. From the results presented in each chapter, it can be concluded that the penalized approaches provide a more accurate and smoother estimation of the functional parameters and have the best forecasting and classification performance. A summary of the main contributions and conclusions for each chapter is given hereafter.

B.1 Chapter 1

Non-penalized and penalized least squares smoothing in terms of B-spline bases have been compared in this chapter to approximate a set of unobserved smooth curves from discrete noisy observations. A simulation study and two applications with real functional data have been developed to study and compare the performance of the three considered smoothers (regression splines, smoothing splines and P-splines) in the FDA context.

In base to these results we can conclude that regression splines and smoothing splines lose control of the smoothness when the number of knots increases. Both penalized approaches get to improve the fit providing mean squared errors with respect to the original smooth sample curves much smaller than the ones given by the non-penalized approach. On the other hand, P-splines provide the lowest approximation errors, have less numerical complexity making easier its computational implementation and are quite insensitive to the choice of knots so that it is sufficient to choose a relatively large number of equally spaced basis knots.

B.2 Chapter 2

Two smoothed FPCA approaches based on P-spline penalties have been proposed in this chapter to control the degree of smoothness of the principal components weight functions estimated from smooth sample curves observed with error. Both approaches are based on B-spline expansion of sample curves and a P-spline penalty that measures the roughness of a function by summing squared d -order differences between adjacent B-spline coefficients. The first smoothed FPCA approach (called FPCA of P-splines) introduces the P-spline penalty in the least squares approximation of the sample curves with B-spline functions (P-splines) and then carries out a non-penalized FPCA on the approximated curves. The second approach approximates the sample curves by non-penalized least squares (regression splines) and then performed a penalized FPCA estimation based on maximizing a penalized sample variance that introduces the P-spline penalty in the orthonormality constraint between principal components.

A simulation study was performed to test the ability of the proposed smoothed FPCA approaches to provide an accurate and smooth estimation of the principal component curves. The results were compared with the estimations provided by non-penalized FPCA of the least squares approximation of sample curves with B-spline basis and regularized FPCA based on penalizing the roughness of the principal component curves by its integrated squared 2-order derivative. From this simulation study it can be concluded that the penalized approaches give much more accurate estimations than non-penalized FPCA. This is because FPCA loses control of the smoothness when the dimension of the B-spline base increases. On the other hand, the smoothed FPCA approaches are quite insensitive to the choice of knots so

that a relatively large number of equally spaced basis knots is a good election for the definition of the B-spline basis. The advantage of the smoothed FPCA approaches based on P-spline penalties respect to the ones based on penalizing the integrated squared d -order derivatives is that they are mathematically simpler because the difference matrix is easier to compute than the matrix of integrals of products of d -order derivatives between B-spline basis functions (see Bhatti and Bracken (2006) for a detailed study on the calculation of integrals involving B-splines). Finally, it can be concluded that FPCA of P-splines is preferable to P-spline SFPCA and regularized FPCA because its computational cost is lower and the approximation errors are slightly smaller.

B.3 Chapter 3

In order to solve the problem of multicollinearity in functional logit regression and to control de smoothness of the functional parameter estimated from noisy smooth sample curves, four different penalized spline (P-spline) estimations of the functional logit model are proposed in this chapter. Let us take into account that the aim of logit model is not only to classify a set of curves into in two groups but mainly to interpret the relationship between the binary response and the functional predictor in terms of the functional parameter. Because of this, our main purpose is to improve the estimation of the functional parameter of a functional logit model, providing in addition a good classification rate.

A P-spline penalty measures the roughness of a curve in terms of differences of order d between coefficients of adjacent B-spline basis functions. The proposed smoothing approaches are based on B-spline expansion of the sample curves and the parameter function, and P-spline estimation of the functional parameter. The difference is in how to introduce the penalty in the model. Three of the considered approaches (Methods II, III and IV) are based on functional principal component logit regression that consists in regressing the binary response on a reduced set of functional principal components. In Method II the P-spline penalty is introduced by performing the functional PCA on the P-spline least squares approximation of the sample curves from discrete observations. Method III introduces the P-spline penalty in the own formulation of functional PCA and the principal components are computed by maximizing a penalized sample variance that introduces a discrete penalty in the orthonormality constraint between the principal components weight func-

tions. In Method IV the P-spline penalty is used in the maximum likelihood estimation of the functional parameter in terms of functional principal components. On the other hand, direct P-spline likelihood estimation in terms of B-spline functions is also considered (Method V).

Two simulation studies were performed to test the ability of the proposed P-spline smoothing approaches to provide an accurate and smooth estimation of the functional parameter and a good classification performance. Leave-one-out cross validation and generalized cross validation are adapted to select the different parameters (smoothing parameter and number of principal components or basis functions) associated with the considered approaches. In the case of the P-spline approximation of the sample curves from equally spaced observations, a relatively large number of equally spaced basis knots is a good choice for the definition of the B-spline basis. The results provided by the different smoothing approaches are compared with the estimations provided by non-penalized FPCLoR on least squares approximation of sample curves with B-spline basis Method I) and by the partial least squares estimation approach for functional linear discriminant analysis with (LDA-FPLS).

From the simulation study it can be concluded that the estimation of the functional parameter given by the P-spline approaches is much smoother than the one given by the non-penalized FPCLoR although in some cases Method IV gives worse results. In fact, Methods I and IV provide non-smooth estimations affected by high variability. The most accurate and smoothest estimations of the parameter function are provided by Methods II and III, based on P-spline estimation of functional PCA with B-spline basis. On the other hand, the estimations given by Method V are less accurate and oversmoothed. In relation to the forecasting ability of the proposed methodologies, Methods II and III provide the least prediction errors followed by Method V that also gives accurate results. The classification performance of all methods is very good, with Methods II, III, and V being the most competitive. On the other hand, the LDA-FPLS approach gives very high classification rates similar to methods II, III and V but its forecasting errors are much higher.

In summary, it can be concluded that the penalized approaches represented by Methods II and III are preferred because they provide the most accurate estimation of the parameter function and have the best forecasting and classification performance, with Method II having lower computational cost.

B.4 Chapter 4

The aim of this chapter is to improve the estimation of the functional parameter associated with the functional linear model for a scalar response when the predictor curves are smooth functions observed with error.

In order to solve the problem of high dimension and multicollinearity in the estimation of the functional linear model, and also to control the degree of smoothness of the estimated functional parameter, two different penalized approaches based on functional partial least squares regression (FPLSR) are developed. The first approach introduces the penalty in the definition of the norm of the PLS component weight functions (Method II). The second one considers a penalized estimation of the covariance between the response and the PLS components (Method III). Discrete and continuous penalties can be used in terms of basis expansions of the sample curves.

Two different criteria based on minimizing the GCVE and the $IMSE_{\beta}$ (criterion 1 and 2, respectively) were adapted to select the different parameters (smoothing parameter and number of PLS components) associated with the considered approaches.

The performance of these penalized FPLS approaches was tested and compared with non-penalized FPLS by using least squares approximation of the sample curves with B-spline basis on a simulation study and an application with chemometric functional data measuring the NIR spectra of gasoline samples. In the simulation study two different schemes were considered so that $R^2 = 0.9$ and $R^2 = 0.7$.

From the simulation study, it can be concluded that the estimation of the functional parameter given by the penalized approaches is much smoother than the one given by the non-penalized FPLS. In fact, it can be said that independently of the model selection criterion and the simulation scheme ($R^2 = 0.9$ or $R^2 = 0.7$), the more accurate estimation of the functional parameter is given by Method II, because the estimations given by Method III are oversmoothed and present more variability. With respect to the forecasting performance, Methods II and III provide similar results, improving both the prediction ability of the non-penalized FPLS approach. The significant differences between the non-penalized and penalized estimations of FPLS are mainly in their capacity to provide an accurate estimation of the functional parameter. On the other hand, using GCV criterion for model selection is a

good option for predicting the response and estimating the parameter function except for Method III where the increment in the number of selected FPLS components worsen the functional parameter estimates slightly.

In the application to the spectroscopic data set of gasoline, the aim was to forecast the octane number from the NIR spectra of 60 gasoline samples, and to get a good estimation of the functional parameter that explains the relationship between the response and the functional predictor. The results of this application corroborates that the penalized FPLS approaches have better forecasting performance and provide smoother estimated parameter than the non-penalized approach, with Method II providing the best results. In both, simulation and application, Method III is which requires the minimum number of predictors.

Summarizing, Method II provides the best estimations of the functional parameter, achieving also a good forecasting performance, and using less predictors than the non-penalized FPLS approach.

B.5 Chapter 5

Different functional classification approaches are applied and compared in this chapter to classify the quality of cookies (good or bad) in terms of the curves of resistance of dough during the kneading process. The aim of this application is to identify those flours that provide good cookies and improve the quality of the manufacturer's products by using only those flours that guarantee the best quality.

Two different classification methods, such as functional logit regression and functional linear discriminant analysis, are considered to classify a set of curves in the two groups defined by a binary response. A third method based on componentwise logit classification is also applied for comparison purposes. The first methodology allows not only to solve the classification problem but also to estimate the relationship between the response variable (quality of cookies) and the predictor variable (resistance of dough during the kneading process). The estimation of the proposed functional classification approaches is affected by several problems such as high dimension and multicollinearity. Estimation based on functional PCA and functional PLS is considered to solve these problems. Smoothed versions of these methodologies based on P-spline approximation of the sample curves with B-spline basis are introduced

in this chapter to solve the problem of lack of smoothness of the estimated functional parameters. Inference on the estimated functional parameters and associated odds ratios is also carried out based on the asymptotic normality of the likelihood estimators.

From the statistical analysis of the results it can be concluded that the proposed functional methodologies (FPCLoR and LDA-FPLS) have a high classification ability with LDA-FPLS being the one that gives the highest area under ROC curve and the minimum misclassification rate. The smoothed versions of both approaches give more accurate and smooth estimations of the functional parameters which facilitates their interpretation. On the other hand, the componentwise logit classifier does not improve the classification ability of the proposed methods because provides the smallest area under ROC curve and the highest misclassification error. The main advantage of this non linear classifier is that provides the highest reduction of dimension because rarely selected more than three or four time points that convey particular information for classification purpose.

Several interpretations of the functional parameter based on odds ratios and principal components are also proposed. To summarize, it can be concluded that good cookies have greater resistance of the dough in the late period and less resistance in the early period. The main features of the curve of resistance of good cookies were also identified by interpreting the first principal component curve.

B.6 Further research

The research line on penalized FDA methodologies is not closed and the authors have in mind to continue working and developing the following ideas:

1. Penalized estimation of functional multinomial response models for nominal and ordinal responses.
2. Comparative study between penalized functional PCR and penalized functional PLS.
3. Formulation and estimation (both non-penalized and penalized approaches) of functional PLS regression when both response and predictor variables are functional.

4. Application of penalized FDA methodologies in different areas of interest as environment, medicine, chemometric, ...



Conclusiones y líneas abiertas

Recordemos que el objetivo principal de esta tesis doctoral es mejorar la estimación de metodologías del ADF para el caso de datos funcionales suaves observados con error. Con objeto de resolver este problema, se han propuesto distintas estimaciones penalizadas con bases de B-splines para el ACP funcional, la regresión logística funcional en componentes principales y el PLS funcional.

El funcionamiento de los métodos penalizados se ha testado mediante estudios de simulación y aplicaciones a datos reales, comparando sus resultados con los obtenidos mediante los correspondientes métodos sin penalización estimados en términos de las representaciones básicas con B-splines de las curvas muestrales y de la función parámetro.

C.1 Capítulo 1

En este capítulo, con objeto de aproximar un conjunto de curvas suaves a partir de observaciones discretas ruidosas, se han comparado distintos suavizados basados en la estimación por mínimos cuadrados penalizados y no penalizados. Se han llevado a cabo un estudio de simulación y varias aplicaciones a datos reales con objeto de comparar las tres formas de suavizado de curvas consideradas (splines de regresión, splines de suavizado y P-splines) en el contexto de ADF.

En base a estos resultados se puede concluir que los splines de regresión y los splines de suavizado pierden el control de la suavidad de la curva a medida que aumenta el número de nodos de la base. Ambas aproximaciones penalizadas (splines de suavizado y P-splines) consiguen mejorar el ajuste de las curvas proporcionando errores cuadráticos medios de aproximación a la curvas muestrales suaves originales más pequeños que los proporcionados con la estimación no penalizada. Por un lado, los P-splines proporcionan los errores de aproximación más bajos, tienen menos complejidad numérica, lo que hace más simple su implementación computacional, y son menos sensibles a la elección de los nodos de la base, siendo suficiente con seleccionar un conjunto relativamente grande de nodos igualmente espaciados.

C.2 Capítulo 2

En este capítulo, con objeto de controlar el grado de suavidad de las funciones peso de las componentes principales estimadas a partir de curvas suaves observadas con error, se proponen dos aproximaciones suaves del análisis en componentes principales funcional (FPCA) basadas en penalizaciones P-spline. Ambas aproximaciones se basan en la representación básica con B-splines de las curvas muestrales y en una penalización P-spline que mide la falta de suavidad de una función sumando los cuadrados de las diferencias de orden d entre los coeficientes adyacentes de los B-splines. La primera aproximación suave del FPCA (llamada FPCA de los P-splines) introduce la penalización P-spline en la aproximación por mínimos cuadrados de las curvas muestrales con funciones B-spline (P-splines) y posteriormente lleva a cabo un FPCA no penalizado sobre las curvas aproximadas. La segunda aproximación considera las curvas muestrales aproximadas mediante mínimos cuadrados no penalizados (splines de regresión) y realiza una estimación penalizada del FPCA basada en la maximización de la varianza muestral penalizada, que introduce la penalización P-spline en la condición de ortonormalidad entre las componentes principales.

Se ha llevado a cabo un estudio de simulación donde se ha testado la capacidad de las estimaciones penalizadas del FPCA para proporcionar una estimación adecuada y suave de las curvas de las componentes principales. Estos resultados se han contrastado con los obtenidos mediante el FPCA no penalizado sobre las curvas muestrales aproximadas por mínimos cuadrados con bases de B-splines y el *regularized* FPCA que penaliza la falta de suavidad

de las curvas de componentes principales mediante la integral del cuadrado de la derivada de orden 2 de dichas curvas. A partir de este estudio de simulación se puede concluir que las estimaciones penalizadas proporcionan estimaciones más adecuadas que el FPCA no penalizado. Esto se debe a que el FPCA no penalizado pierde el control de la suavidad cuando la dimensión de la base de B-splines aumenta. Por otro lado, las aproximaciones suaves del FPCA son poco sensibles a la selección de los nodos básicos, de modo que bastaría con utilizar un número considerablemente grande de nodos básicos equidistantes para definir la base de B-splines. La ventaja de las estimaciones penalizadas del FPCA basadas en la penalización P-spline respecto a la basada en la penalización continua es que las primeras son matemáticamente más simples, ya que es más fácil calcular la matriz de diferencias que la matriz de las integrales de los productos de las derivadas de orden 2 entre funciones básicas de B-spline (ver Bhatti and Bracken (2006) para un estudio detallado sobre el cálculo de las integrales con B-splines). Finalmente, se puede concluir que el FPCA sobre los P-splines es preferible al FPCA suavizado con penalización P-spline y al *regularized* FPCA, ya que su coste computacional es menor y los errores de aproximación de las curvas son ligeramente inferiores.

C.3 Capítulo 3

Con objeto de resolver el problema de la multicolinealidad en el modelo de regresión logística funcional y controlar la suavidad del parámetro funcional estimado a partir de curvas muestrales suaves observadas con error, en este capítulo se proponen cuatro estimaciones P-spline penalizadas del modelo logit funcional. Hay que tener en cuenta que el objetivo principal del modelo logit no es tanto clasificar un conjunto de curvas en dos grupos, sino interpretar la relación entre la variable respuesta y el predictor funcional en términos de la función parámetro. Por ello, el principal propósito de este trabajo es mejorar la estimación de la función parámetro de un modelo logit funcional, proporcionando al mismo tiempo una tasa de clasificaciones correctas adecuada.

Una penalización P-spline mide la falta de suavidad de una curva en términos de diferencias de orden d entre coeficientes básicos de B-splines adyacentes. Los modelos penalizados propuestos se basan en la expansión básica con B-splines de las curvas muestrales y la función parámetro, y una estimación P-spline de la función parámetro. La diferencia radica en cómo se

introduce la penalización en el modelo. Tres de las aproximaciones propuestas (Métodos II, III y IV) se basan en la regresión en componentes principales funcional, es decir, la regresión de una variable respuesta binaria sobre un conjunto reducido de componentes principales funcionales. En el Método II, la penalización P-spline se introduce considerando un FPCA sobre la aproximación P-spline de las observaciones discretas de las curvas muestrales (P-splines). El Método III introduce la penalización en la propia formulación del PCA, de modo que las componentes principales se obtienen maximizando una varianza penalizada que introduce una penalización discreta en la condición de ortonormalidad entre las funciones peso de las componentes principales. En el Método IV, la penalización se utiliza en la estimación por máxima verosimilitud del parámetro funcional en términos de componentes principales. Por otro lado, se considera una penalización P-spline directa de la verosimilitud en términos de bases de B-splines (Método V).

Con objeto de testar la capacidad de los métodos propuestos tanto para predecir, como para obtener una estimación adecuada y suave de la función parámetro, se han desarrollado dos estudios de simulación. Los métodos de validación cruzada leave-one-out y validación cruzada generalizada se han adaptado para seleccionar los parámetros de los modelos propuestos (parámetro de suavizado y número de componentes principales o funciones básicas). Para la aproximación P-spline de las curvas muestrales a partir de observaciones equidistantes basta con seleccionar un número considerablemente grande de nodos para la definición de la base de B-splines. Los resultados obtenidos mediante las metodologías penalizadas se han comparado con las estimaciones proporcionadas por la regresión logística funcional en componentes principales (FPCLoR) no penalizada sobre las curvas muestrales aproximadas con bases de B-splines (Método I) y con el análisis discriminante lineal basado en el PLS funcional (LDA-FPLS).

A partir del estudio de simulación se puede concluir que la estimación del parámetro funcional proporcionada por los métodos basados en la penalización P-spline es más adecuada que la proporcionada por FPCLoR no penalizada, aunque hay ocasiones en las que el Método IV no da buenas estimaciones. De hecho, los Métodos I y IV proporcionan unas estimaciones poco suaves y con gran variabilidad. Las estimaciones más adecuadas y suaves de la función parámetro se han obtenido con los Métodos II y III, ambos basados en la estimación P-spline del FPCA con bases de B-splines. Por otro lado, las estimaciones proporcionadas por el Método V son menos adecuadas

y tienden a ser demasiado suaves. En relación a la capacidad predictora de las metodologías expuestas, los Métodos II y III son los que proporcionan menores errores de predicción, seguidos del Método V que también proporciona buenos resultados. La capacidad de clasificación de todos los métodos es bastante buena en general, siendo los Métodos II, III y V los más competitivos en este sentido. Por otro lado, el LDA-FPLS proporciona una tasa de clasificación elevada y similar a la de los Métodos II, III y V, pero sus errores de predicción son mayores.

Resumiendo, se puede concluir que las estimaciones penalizadas basadas en los Métodos II y III son las más adecuadas, en el sentido de que proporcionan las mejores estimaciones de la función parámetro y tienen mejor capacidad de predicción y clasificación, siendo el Método II el que supone un menor coste computacional.

C.4 Capítulo 4

El objetivo de este capítulo es mejorar la estimación de la función parámetro del modelo lineal funcional para una variable respuesta escalar cuando los predictores son curvas suaves observadas con error.

Con objeto de resolver el problema de alta dimensión y multicolinealidad en la estimación del modelo lineal funcional, y al mismo tiempo controlar el grado de suavidad de la función parámetro estimada, se proponen dos versiones penalizadas distintas basadas en la regresión PLS funcional (FPLS). La primera aproximación introduce la penalización en la definición de la norma de las funciones peso de las componentes PLS (Método II). La segunda considera una estimación penalizada de la covarianza entre la respuesta y las componentes PLS (Método III). Se puede utilizar tanto la penalización discreta como la continua, ambas en términos de la representación básica de las curvas muestrales.

Para seleccionar los distintos parámetros asociados a las dos aproximaciones propuestas (parámetros de suavizado y número de componentes PLS), se han adaptado dos criterios diferentes basados en minimizar los errores de validación cruzada generalizada y los errores cuadráticos medios integrados de la función parámetro (criterios 1 y 2, respectivamente).

El comportamiento de las aproximaciones penalizadas de la regresión PLS funcional propuestas en este capítulo se ha testado mediante un estudio de

simulación y una aplicación con datos reales basados en el espectro NIR de muestras de gasolina, comparando tales resultados con los del FPLS no penalizado. En el estudio de simulación se consideraron dos esquemas distintos, uno para $R^2 = 0.7$ y otro para $R^2 = 0.9$.

A partir del estudio de simulación podemos concluir que los métodos penalizados proporcionan una función parámetro más suave y adecuada que el PLSF no penalizado. Además, con independencia del criterio de selección de parámetros empleado y del esquema de simulación seguido ($R^2 = 0.7$ o $R^2 = 0.9$), la mejor estimación de la función parámetro viene dada por el Método II, ya que las estimaciones proporcionadas por el Método III son demasiado suaves y presentan mucha variabilidad. Respecto a la capacidad de predicción, los Métodos II y III presentan resultados muy similares, mejorando ambos las predicciones dadas por el FPLS no penalizado. No obstante, podemos decir que las diferencias más significativas entre las aproximaciones penalizadas y no penalizada radican en la estimación de la función parámetro. Por otro lado, es una buena opción utilizar el criterio de validación cruzada generalizada en la selección de los modelos para predecir y estimar la función parámetro, excepto para el Método III, donde el incremento en el número de componentes PLS seleccionadas empeora ligeramente la estimación de la función parámetro.

En la aplicación a los datos sobre espectros de la gasolina, el objetivo fue predecir el número de octanos a partir del espectro NIR de 60 muestras de gasolina y conseguir una buena estimación de la función parámetro que explique la relación entre la variable respuesta y el predictor funcional. Los resultados de esta aplicación corroboran que las aproximaciones penalizadas del FPLS tienen mejor capacidad predictora y consiguen una función parámetro más suave que la aproximación no penalizada, siendo el Método II el que proporciona los mejores resultados. Tanto en la simulación como en la aplicación, el Método III es el que requiere el mínimo número de predictores.

Resumiendo, el Método II proporciona las mejores estimaciones de la función parámetro, consiguiendo al mismo tiempo una buena capacidad predictora y utilizando menos variables en el modelo (componentes PLS) que el FPLS no penalizado.

C.5 Capítulo 5

En este capítulo se proponen y comparan distintas metodologías para la clasificación funcional con objeto de clasificar la calidad de las galletas (buenas o malas) en términos de las curvas de resistencia de la masa de las mismas durante el proceso de horneado. El objetivo de esta aplicación es identificar aquellas harinas que dan lugar a galletas buenas, mejorando así la calidad de los productos de la manufacturera DANONE usando sólo harinas que garanticen la mejor calidad.

Con objeto de clasificar un conjunto de curvas en dos grupos definidos por una variable respuesta binaria, se han considerado dos métodos de clasificación funcional como son la regresión logística funcional y el análisis discriminante lineal funcional. A modo comparativo se propone otro método basado en la clasificación logit componente a componente (*componentwise classification*). El primer método no sólo permite clasificar las curvas en dos grupos, sino también estimar la relación entre la variable respuesta (calidad de las galletas) y el predictor funcional (resistencia de la masa durante el proceso de horneado). La estimación de los distintos métodos de clasificación funcional propuestos en este capítulo están afectados por la alta dimensión de los datos y la multicolinealidad. Para resolver ambos problemas se considera una estimación de los métodos mencionados previamente basada en FPCA y FPLS. Con objeto de resolver el problema de falta de suavidad de los parámetros funcionales estimados, se han propuesto versiones suavizadas de estas metodologías basadas en la aproximación P-spline de la curvas muestrales con bases de B-splines. Basándonos en la normalidad asintótica de los estimadores de máxima verosimilitud, se lleva a cabo inferencia sobre los parámetros funcionales estimados y los cocientes de ventajas asociados.

A partir del análisis estadístico de los resultados se puede concluir que las metodologías funcionales propuestas (FPCLoR y LDA-FPLS) presentan una elevada capacidad de clasificación, siendo LDA-FPLS el que proporciona las mayores áreas bajo la curva ROC y las tasas de clasificaciones incorrectas más bajas. Las versiones penalizadas de ambas aproximaciones proporcionan una estimación más suave y adecuada de las funciones parámetro, lo cual facilita su interpretación. Por otro lado, el clasificador logit componente a componente no mejora la capacidad de clasificación de los métodos propuestos, dando lugar a áreas bajo la curva ROC más pequeñas y tasas de clasificaciones incorrectas más elevadas. No obstante, la principal ventaja de

este clasificador funcional es que consigue reducir bastante de la dimensión del problema, ya que rara vez selecciona más de tres o cuatro instantes de tiempo que contienen información importante para la clasificación.

Se proponen varias interpretaciones de la función parámetro en base a los cocientes de ventajas y las componentes principales. Resumiendo, se puede concluir que las galletas de buena calidad presentan una resistencia de la masa mayor al final de la fase de horneado y menos resistencia al principio. Las principales características de las curvas de resistencia de las galletas buenas también se han identificado interpretando la curva de la primera componente principal.

C.6 Líneas abiertas

La línea de investigación sobre metodologías del ADF penalizadas no termina con esta tesis doctoral. Los autores tienen en mente continuar trabajando y desarrollando las siguientes ideas:

1. Estimación penalizada de los modelos de respuesta multinomial funcional para respuestas nominales y ordinales.
2. Estudios comparativos entre PCR funcional penalizado y PLS funcional penalizado.
3. Formulación y estimación (aproximaciones penalizadas y no penalizadas) de la regresión PLS cuando la respuesta y las variables predictoras son funcionales.
4. Aplicación de las metodologías penalizadas del ADF en distintas áreas de interés como medio ambiente, medicina, quimiometría, ...

Summary

A functional variable is characterized because its observations are functions that in the majority of cases represent the evolution of a scalar variable in time (realizations of a stochastic process). This is the case of environmental variables such as temperature or contamination level observed daily in a period of time, economic variables such as stock price evolution or medical variables such as stress level. In other areas of application the argument of the observed functions is a different magnitude such as spatial location, wavelength or probability. In many chemometric applications, observations of the NIR spectrum at a fine grid of wavelengths are available.

Functional data analysis (FDA) is an statistical topic of active research devoted to solve problems related with the statistical modeling and prediction of functional data. An overview of the basic methods of FDA, computational aspects related with their practical application and important real data modeling can be seen in the pioneers books by Ramsay and Silverman (2005, 2002); Ramsay et al. (2009). A detailed study on nonparametric FDA methodologies was developed in Ferraty and Vieu (2006). Statistical inference related with some FDA methods was recently studied in Horvath and Kokoszka (2012).

Early work on FDA was developed in the framework of continuous-time stochastic processes and were devoted to the generalization of reduction dimension techniques such as principal component analysis (PCA) to the functional case (Deville, 1974). Later, statistical researching on FDA focused on the formulation and estimation of different functional regression models.

The functional linear model to estimate a scalar response from a functional predictor was one of the first regression models extended to the case of functional data (Cardot et al., 1999, 2003). The case where the predictor is a vector or scalar and the response is functional was studied by Chiou et al. (2004). Functional analysis of variance was introduced to model the mean of a functional response in terms of a categorical variable (Cuevas et al., 2002, 2004). On the other hand, functional linear models where both predictor and response variables are functional were studied by Yao et al. (2005b) and Ocaña et al. (2008). Principal component prediction models, that can be seen as a particular case of these linear models, were first introduced to forecast a continuous time stochastic process on a future interval from its recent past (Aguilera et al., 1997, 1999). On the other hand, generalized linear models were also extended for the case of a functional predictor (James, 2002; Müller, 2005). A particular case of functional generalized model is the functional logit regression model whose aim is to predict a binary random variable from a functional predictor (Ratcliffe et al., 2002; Escabias et al., 2004; Aguilera et al., 2008b).

Direct estimation of the functional parameter associated with a functional regression model is an ill-posed problem due to the infinite dimension of the functional variable. On the other hand, sample curves are usually observed in a finite set of sampling points that could be unequally spaced and different among the sample units. Because of this the first step in FDA is to reconstruct the true functional form of each sample curve from a finite set of discrete observations. Approximation techniques such as interpolation or projection in a finite-dimensional space generated by basis functions were applied from the beginning to solve these problems. This way, the estimation of a functional regression model is reduced to the estimation of an equivalent multivariate regression model with high correlation between the predictor variables.

Regression on a set of uncorrelated random variables is usually used in literature to provide an accurate estimation of the parameters associated with a regression model. Functional PCA was used to reduce the dimension and solving the multicollinearity problem in many functional regression models. Principal components are uncorrelated generalized linear combinations of the functional predictor with maximum variance. Because of this the main criticism about principal component regression is that the regressors are computed without taking into account the response variable. To solve this problem, functional partial least squares was extended to the functional

case by computing a set of uncorrelated generalized linear combinations of the predictor variable having maximum covariance with the response variable (Preda and Saporta, 2005b).

In many applications the data are smooth functions observed with error. In this case least squares approximation with B-spline bases is usually used to estimate the basis coefficients of a basis expansion of the unobserved smooth sample functions. The problem is that the approximated sample curves (regression splines) do not control the degree of smoothness. As a consequence, the estimated principal components and functional parameters associated with functional regression models are difficult to interpret because they have a lot of variability and lack of smoothness.

The general objective of this thesis is to improve the estimation of FDA methodologies in the case of smooth functional data observed with error. In order to solve this problem, different approaches based on penalized estimation with B-spline basis expansions of sample curves are proposed. This general objective is achieved through five specific objectives:

1. Review and comparison of existing methods for the approximation of smooth curves with B-splines bases.
2. Improve the estimation of functional PCA by introducing different penalized spline approaches.
3. Develop different penalized approaches for estimating the functional logit model based on penalized spline estimation of functional PCA.
4. Propose different penalized estimation approaches in functional PLS regression.
5. Develop an application of the proposed penalized estimation methodologies to improve the quality in food industry.

According with the specific objectives, the thesis is divided into five chapters with the methodology and results related with each one. The contents of each chapter have been included in different research papers actually submitted or accepted for publication in different JCR journals. In addition to the methodological contributions in each chapter, the proposed penalized FDA methods were applied on simulated and real data by developing own code

with the free statistical software R (<http://www.r-project.org>). A brief description of the main libraries and functions used in this thesis can be seen in Appendix A at the end of this memory.

D.1 Chapter 1

The main purpose of this chapter is to review and compare three different approaches for approximating smooth sample curves observed with error in terms of B-spline bases: regression splines (non-penalized least squares approximation), smoothing splines (continuous roughness penalty based on the integrated squared d -order derivative of each sample curve) and P-splines (discrete roughness penalty based on d -order differences between coefficients of adjacent B-splines). The performance of these spline smoothing approaches is studied via a simulation study and several applications with real data. Cross-validation and generalized cross-validation are adapted to select a common smoothing parameter for all sample curves with the roughness penalty approaches.

The approximation of smooth noisy functions with B-spline bases is used in the estimation of a wide variety of FDA methodologies. This justifies the importance of a comparison among the main smoothing approaches in terms of B-splines and to draw conclusions that allow the researchers and practitioners to use the most powerful tool in each case.

Let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of size n of a functional variable X , whose observations are independent and equally distributed realizations of a second order stochastic process $X = \{X(t) : t \in T\}$, continuous in quadratic mean, whose sample functions belong to the Hilbert space $L^2(T)$ of square integrable functions with the usual scalar product given by

$$\langle f, g \rangle = \int_T f(t) g(t) dt, \quad \forall f, g \in L^2(T).$$

The sample paths are assumed to belong to a finite-dimension space generated by a basis $\{\phi_1(t), \dots, \phi_p(t)\}$ so that they are expressed as

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t) = a_i' \phi_j(t), \quad i = 1, \dots, n,$$

where $a_i = (a_{i1}, \dots, a_{ip})'$ is the vector of basis coefficients of the i -th sample path.

In this chapter, the vectors of basis coefficients of the sample paths are estimated by different methodologies, such as regression splines (without roughness penalty), smoothing splines (continuous penalty) and P-splines (discrete penalty).

- Regression splines: $\hat{a}_i = (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i$, with x_i being the values of the sample paths at the observation knots, and $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$.
- Smoothing splines: $\hat{a}_i = (\Phi_i' \Phi_i + \lambda R_d)^{-1} \Phi_i' x_i$, where λ is the smoothing parameter, and R_d is a matrix whose elements are the integrals of products of d-order derivatives between B-spline basis functions.
- P-splines: $\hat{a}_i = (\Phi_i' \Phi_i + \lambda P_d)^{-1} \Phi_i' x_i$, where $P_d = (\Delta^d)' \Delta^d$, with Δ^d being a matrix whose elements are the d-order differences between coefficients of adjacent B-splines.

D.2 Chapter 2

Functional principal component analysis (FPCA) is a dimension reduction technique that explains the dependence structure of a functional data set in terms of uncorrelated variables. In many applications the data are a set of smooth functions observed with error. In these cases the principal components are difficult to interpret because the estimated weight functions have a lot of variability and lack of smoothness. The most common way to solve this problem is based on penalizing the roughness of a function by its integrated squared d-order derivative.

In this chapter, two alternative forms of penalized FPCA based on B-spline basis expansions of sample curves and a simpler P-spline penalty are proposed. The main difference between both smoothed FPCA approaches is that the first uses the P-spline penalty in the least squares approximation of the sample curves in terms of a B-spline basis meanwhile the second introduces the P-spline penalty in the orthonormality constraint of the algorithm that computes the principal components. Leave-one-out cross-validation is adapted to select the smoothing parameter for these two smoothed FPCA approaches. A simulation study and an application with chemometric functional data are developed to test the performance of the proposed penalized approaches and to compare the results with non-penalized FPCA and regularized FPCA.

In the same conditions of Chapter 1, let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of functions that are the sample information related to a functional variable X . In general the j -th principal component scores is expressed in terms of the weight functions or loadings as follows

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

where the associated weight functions are the eigenfunctions of the sample covariance operator. That is, the solutions to the eigenequation

$$\mathcal{C}(f_j)(t) = \int_T C(t, s) f_j(s) ds = \lambda_j f_j(t).$$

Considering the basis representation of the sample curves, the weight functions can be expressed in terms of the same basis so that

$$f_j(t) = \sum_{k=1}^p b_{jk} \phi_k(t) = \phi(t)' b_j,$$

with $b_j = (b_{j1}, \dots, b_{jp})'$. The estimation of the vector of basis coefficients of the weight functions depends on the type of FPCA to be used.

- Non-penalized FPCA is equivalent to the multivariate PCA of the matrix $A\Psi^{\frac{1}{2}}$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines and $\Psi^{\frac{1}{2}}$ being the squared root of the matrix of inner products between basis functions (Ocaña et al., 2007).
- Functional PCA of P-splines is equivalent to the multivariate PCA of the matrix $A\Psi^{\frac{1}{2}}$, with A being the matrix of basis coefficients of the sample paths approximated by P-splines.

In order to compute the optimal value of the smoothing parameter λ , two selection criteria are considered and compared in this chapter: leave-one-out cross validation (CV) and generalized cross validation (GCV).

- P-spline smoothed FPCA is reduced to multivariate PCA of the matrix $A\Psi L^{-1}$, where A is the matrix of basis coefficients of the sample paths approximated by regression splines, L^{-1} is the inverse of a lower triangular matrix obtained by the Choleski factorization of $LL' = \Psi + \lambda P_d$, with λ being the smoothing parameter and P_d being the discrete penalty matrix defined in Chapter 1.

In order to select the smoothing parameter in P-spline smoothed FPCA, leave one out cross-validation (CV) has been adapted by considering the quadratic distances in terms of basis representations. It consists of selecting the value of λ that minimizes

$$CV(\lambda) = \frac{1}{p} \sum_{q=1}^p CV_q(\lambda),$$

where

$$CV_q(\lambda) = \frac{1}{n} \sum_{i=1}^n \|x_i - x_i^{q(-i)}\|^2,$$

with $x_i^{q(-i)} = \sum_{l=1}^q \xi_{il}^{(-i)} f_l^{(-i)}$ being the reconstruction of the sample curve x_i in terms of the first q principal components estimated from the sample of size $n - 1$ that includes all sample curves except x_i .

In terms of basis expansions these quadratic distances are given by

$$\|x_i - x_i^{q(-i)}\|^2 = \int_T [x_i(t) - x_i^{q(-i)}(t)]^2 dt = d_i' \Psi d_i,$$

where $d_i = (d_{i1}, \dots, d_{ip})'$, with $d_{ij} = a_{ij} - \sum_{l=1}^q \xi_{il}^{(-i)} b_{lj}^{(-i)}$, and Ψ is the matrix of inner products between basis functions.

D.3 Chapter 3

This chapter is devoted to improve the estimation of the functional logit model. The problem of multicollinearity associated with the estimation of this 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 non-smooth and then difficult to interpret. To solve this problem different penalized spline estimations of the functional logit model are proposed in this chapter. All of them are based on smoothed functional PCA and/or a discrete P-spline penalty in the log-likelihood criterion in terms of B-spline expansions of the sample curves and the functional parameter.

In the context of functional principal component logit regression, three different versions of penalized estimation approaches based on smoothed FPCA are introduced. On the one hand, FPCA of P-spline approximation of sample curves (Method II) is performed. On the other hand, a discrete P-spline

penalty, that penalizes the roughness of the principal component weight functions, is included in the own formulation of FPCA (Method III). The third smoothed approach is carried out by introducing the penalty in the likelihood estimation of the functional parameter in terms of a reduced set of functional principal components (Method IV). Moreover, direct P-spline likelihood estimation in terms of B-spline functions is also considered (Method V). The ability of these smoothing approaches to provide an accurate estimation of the functional parameter and their classification performance with respect to non-penalized functional PCA are evaluated via simulation and application to real data. Leave-one-out cross-validation and generalized cross-validation are adapted to select the smoothing parameter and the number of principal components or basis functions associated with the considered approaches.

In the same conditions of Chapter 1, let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of functions that are the sample information related to a functional variable X and $\{y_i, i = 1, \dots, n\}$ be a random sample of Y associated with them. That is, $y_i \in \{0, 1\}$, $i = 1, \dots, n$. The functional logistic regression model is given by

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

with the logit transformations expressed as

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

Let us consider the basis representations of the sample paths and the parameter function

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad \beta(t) = \sum_{k=1}^p \beta_k \phi_k(t),$$

with $\beta = (\beta_1, \dots, \beta_p)'$ being the vector of basis coefficients of $\beta(t)$. In this context, different ways to estimated de vector of basis coefficients β are developed in this chapter.

1. FLoM in terms of principal components (FPCLoR).

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

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

where $\Gamma = (\xi_{ij})_{n \times p}$ is a matrix of functional principal components, γ is the vector of coefficients of the model and α is the intercept.

An accurate estimation of the functional parameter can be obtained by considering only a set of q optimum principal components as predictor variables, so that $\Gamma = (\xi_{ij})_{n \times q}$ ($q < p$).

Then, the vector β of basis coefficients is given by $\beta = F\gamma$, where the way of estimating F depends on the kind of FPCA used to estimate the functional model and the kind of likelihood estimation (penalized or non-penalized).

- Method I: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines. In this method γ is estimated by maximum likelihood without penalty.
- Method II: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$, with A being the matrix of basis coefficients of the sample paths estimated by P-splines. In this method γ is estimated by maximum likelihood without penalty.
- Method III: $\hat{\beta} = (L^{-1})' G \hat{\gamma}$, where G is the matrix of eigenvectors of the sample covariance matrix of $A\Psi(L^{-1})'$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines and L defined in Chapter 2. In this method γ is estimated by the maximum likelihood criterion without penalty.
- Method IV: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, where G is the matrix whose columns are the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines. In this method γ is estimated by penalized maximum likelihood so that

$$\mathcal{L}^*(\lambda, \gamma) = \mathcal{L}(\gamma) - \frac{\lambda}{2} \gamma' P_d \gamma,$$

with $\mathcal{L}(\gamma)$ being the log-likelihood of FPCLoR model, λ the smoothing parameter and P_d the discrete penalty matrix defined in Chapter 1.

2. FLoM via penalized log-likelihood.

Let us consider the basis representation of the sample paths and the parameter function. Then, the logit transformations in matrix form are given by

$$L = X\beta,$$

where $L = (l_1, \dots, l_n)$ is the vector of logit transformations, $X = (\mathbf{1}|A\Psi)$, with $\mathbf{1} = (1, \dots, 1)'$ an n -dimensional vector of ones and $\beta = (\beta_1, \dots, \beta_p)'$ the vector of basis coefficients of $\beta(t)$.

Method V: the parameter function is estimated by penalized maximum likelihood and without using principal components. Then, the penalized log-likelihood of the FLoM is given by

$$\mathcal{L}^*(\lambda, \beta) = \mathcal{L}(\beta) - \frac{\lambda}{2}\beta'P_d\beta,$$

with $\mathcal{L}(\gamma)$ being the log-likelihood of FLoM. In this case, the Newton-Raphson solution for the penalized likelihood estimators would be

$$\beta^{(t)} = \beta^{(t-1)} + [X' \text{Diag}(\pi_i^{(t-1)}(1 - \pi_i^{(t-1)}))X + \lambda P_d]^{-1} X'(y - \pi_i^{(t-1)}),$$

with $X = (\mathbf{1}|A\Psi)$.

D.4 Chapter 4

The main problems associated with the functional linear model for a scalar response in terms of smooth curves observed with error, are high dimension, multicollinearity and the lack of smoothness in the functional parameter estimation. In order to solve the three problems at the same time, two different penalized approaches based on partial least squares regression are developed. The main difference between the two proposed approaches is the way in which the penalty is introduced. The first approach introduces the penalty in the definition of the norm of the PLS component weight functions (Method II). The second one considers a penalized estimation of the covariance between the response and the PLS components (Method III). Discrete and continuous penalty are considered in terms of basis expansions of the sample curves. The selection of the optimum number of PLS components and the smoothing parameter is carried out by two different criterion based on GCV errors and the integrated mean squared errors of the parameter function.

In order to test the performance of the proposed penalized FPLS approaches and to compare the results with non-penalized FPLS, a simulation study and an application with chemometric functional data are developed.

In the same conditions of Chapter 1, let $\{x_i(t) : t \in T, i = 1, \dots, n\}$ be a sample of functions that are the sample information related to a functional variable X and $\{y_i : i = 1, \dots, n\}$ be a random sample of Y associated with them. The FLM is then expressed as

$$y_i = \beta_0 + \int_T x_i(t)\beta(t)dt + \varepsilon_i,$$

with $\{\varepsilon_i : i = 1, \dots, n\}$ being independent and centered random errors.

In order to get an accurate estimation of the parameter function, the FLM model is considered in terms of PLS components so that

$$Y = \mathbf{1}\gamma_0 + T\gamma,$$

where T is the matrix comprising the columns of the PLS components and γ is the vector of the regression coefficients of Y on T .

An accurate estimation of the functional parameter can be obtained by considering only a set of q optimum PLS components as predictor variables so that $T = (t_{ij})_{n \times q}$ ($q < p$). The matrix of PLS components T is estimated by three different versions of FPLS

Let us consider the basis representation of the sample paths and the parameter function

$$x_i(t) = \sum_{j=1}^p a_{ij}\phi_j(t), \quad \beta(t) = \sum_{k=1}^p \beta_k\phi_k(t).$$

Then, the following versions of FPLS regression are considered:

- Non-penalized FPLS.

The functional PLS regression (FPLS) of a real random response Y in terms of a functional predictor $X = \{X(t) : t \in T\}$ is an iterative procedure so that the h -th PLS component is given by

$$t_h = \int_T X_{h-1}(t) w_h(t) dt,$$

with $w_h(t)$ being the weight function obtained by solving the following problem:

$$\begin{aligned} \max_w \quad & Cov^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right), \\ & \|w\|^2 = 1 \end{aligned}$$

where $X_0(t) = X(t)$, $\forall t \in T$ and $Y_0 = Y$. The h -th PLS step is concluded with the linear regression of $X_{h-1}(t)$ and Y_{h-1} on t_h so that

$$\begin{aligned} X_h(t) &= X_{h-1}(t) - p_h(t) t_h, \quad t \in T \\ Y_h &= Y_{h-1} - c_h t_h, \end{aligned}$$

where $p_h(t) = (\mathbb{E}(X_{h-1}(t) t_h) / \mathbb{E}(t_h^2))$ and $c_h = (\mathbb{E}(Y_{h-1} t_h) / \mathbb{E}(t_h^2))$.

Non-penalized FPLS is reduced to a multivariate PLS of Y on the matrix $A\Psi^{1/2}$ so that $T = A\Psi^{1/2}V$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines and V being the matrix comprising the columns of the eigenvectors associated with the estimated PLS components.

The vector of basis coefficients of the functional parameter β is estimated by $\hat{\beta} = (\Psi^{-1/2})'V\hat{\gamma}$.

- FPLS by penalizing the norm.

In this case, the h -th penalized PLS component is achieved by the following maximization problem

$$\max_w \frac{Cov^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right)}{\langle w, w \rangle + \lambda PEN_d(w)},$$

where $\langle w, w \rangle$ is the classical scalar product, $PEN_d(w)$ is a general penalty defined in Chapter 1 and λ is the smoothing parameter.

This version of Penalized FPLS is reduced to a multivariate PLS of Y on the matrix $A\Psi(L^{-1})'$ so that $T = A\Psi(L^{-1})'V$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines, V being the matrix comprising the columns of the eigenvectors associated with the estimated PLS components and L a matrix defined in Chapter 2.

The vector of basis coefficients of the functional parameter β is estimated by $\hat{\beta} = (L^{-1})'V\hat{\gamma}$.

- FPLS by penalizing the covariance.

In this case, the h -th penalized PLS component is achieved by the following maximization problem

$$\max_w \frac{\text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right) - \lambda \text{PEN}_d(w)}{\langle w, w \rangle},$$

where $\langle w, w \rangle$ is the classical scalar product, $\text{PEN}_d(w)$ is a general penalty defined in Chapter 1 and λ is the smoothing parameter.

In this version of Penalized FPLS the matrix of penalized PLS components is given by $T = A\Psi(\Psi^{-1/2})'V$, with A being the matrix of basis coefficients of the sample paths approximated by regression splines and V being the matrix comprising the columns of the eigenvectors associated with the estimated PLS components.

In this case, the vector of basis coefficients of the functional parameter β is estimated by $\hat{\beta} = (\Psi^{-1/2})'V\hat{\gamma}$.

D.5 Chapter 5

The aim of this chapter is to improve the quality of cookies production by classifying them as good or bad from the curves of resistance of dough observed during the kneading process. As the predictor variable is functional, functional classification methodologies such as functional logit regression and functional discriminant analysis are considered. A P-spline approximation of the sample curves is proposed to improve the classification ability of these models and to suitably estimate the relationship between the quality of cookies and the resistance of dough. Inference results on the functional parameters and related odds ratios are obtained using the asymptotic normality of the maximum likelihood estimators under the classical regularity conditions. Finally, the classification results are compared with alternative functional data analysis approaches such as componentwise classification on the logit regression model.

Let us consider the classification problem of a sample of functional observations $\{x_i(t) : t \in T; i = 1, \dots, n\}$ according to a related binary response $Y \in \{0, 1\}$ whose observations are denoted by $\{y_i : i = 1, \dots, n\}$.

1. Penalized FPCLoR (Method II Chapter 3)

Let us remember the FLoM defined in Chapter 3 and the associated logit transformations expressed as

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

- Functional parameter interpretation

$$\exp(l_i^* - l_i) = \exp \left(K \int_{t_0}^{t_0+h} \beta(t) dt \right),$$

where l_i^* is the logit transformation for the i -th functional observation constantly increased in the period $[t_0, t_0+h]$, and K is a positive constant. This is an odds ratio, so that the odds of outcome $Y = 1$ is multiplied by this amount when the value of the functional observation is constantly increased in K units in a fixed interval $[t_0, t_0+h]$.

- Inference on the functional parameter

Let $\hat{I} = \int_{t_0}^{t_0+h} \hat{\beta}(t) dt$ be the maximum likelihood estimator for $I = \int_{t_0}^{t_0+h} \beta(t) dt$, with $\hat{\beta}(t)$ being the maximum likelihood estimator of $\beta(t)$.

A $(1 - \alpha)$ confidence interval for the odds ratio is given by

$$\exp(\hat{I} \pm \hat{\sigma}(\hat{I})z_{\alpha/2}).$$

A $(1 - \alpha)$ pointwise confidence interval for $\beta(t)$ is given by

$$\hat{\beta}(t) \pm \hat{\sigma}(\hat{\beta}(t))z_{\alpha/2}.$$

Let us consider the basis representation of the sample paths and the functional parameter. Then, the vector of basis coefficients of $\beta(t)$ is estimated (Method II Chapter 3) as $\hat{\beta} = F\hat{\gamma}$. Then,

$$Var[\hat{I}] = \int_{t_0}^{t_0+h} \int_{t_0}^{t_0+h} Cov(\hat{\beta}(t), \hat{\beta}(s)) dt ds$$

can be approximated by

$$Var[\hat{I}] = \int \int \phi'(t) FCov(\hat{\gamma}) F' \phi(s) dt ds.$$

2. Penalized LDA-FPLS

Linear discriminant analysis (LDA) in the functional data context aims to find linear combinations $\int_T X(t) \beta(t) dt$, $\beta \in L_2(T)$, so that the variance between classes is maximized with respect to the total variance

$$\max_{\beta} \frac{\text{Var}(E[X(t)|Y])}{\text{Var}(X(t))}.$$

The discriminant function is the coefficient function of the functional linear regression of Y on $\{X(t) : t \in T\}$, where Y is recoded as follows

$$\begin{aligned} Y &= -\sqrt{p_0/p_1} \quad \text{if } Y=1 \\ Y &= \sqrt{p_1/p_0} \quad \text{if } Y=0, \end{aligned}$$

with $p_0 = P[Y = 0]$ and $p_1 = P[Y = 1]$.

Because of the equivalence between linear discriminant analysis and linear regression, the problem of high dimension is solved by using a set of FPLS components as predictor variables. Let us consider that the sample curves are affected by noise. Because of this, the estimated discriminant function is non-smooth and a smoothed estimation is proposed. As in FPCLoR, this estimation is based on the previous P-spline approximation of the sample curves with B-spline functions.

3. Componentwise classification

The componentwise classification approach consists of determining a relatively small number of points $\{t_1^*, \dots, t_p^* \in T\}$ that have important leverage for classification and applying a standard classification method on the vector $(X(t_1^*), \dots, X(t_p^*))$. A detailed study on the theoretical properties of the method and its behavior for different classifiers can be seen in Delaigle et al. (2012). In this chapter we will illustrate this approach for the classifier based on the logit regression model. The resulting logit-based componentwise classification approach has two main steps based on selecting the value of p and the position of the set of optimum time knots adaptively.

Resumen

Una variable funcional se caracteriza porque sus observaciones son funciones que en la mayoría de los casos representan la evolución de una variable escalar en el tiempo (realizaciones de un proceso estocástico). Este es el caso de variables medioambientales tales como son la temperatura o el nivel de contaminación observados diariamente durante un periodo de tiempo, variables económicas tales como la evolución de las cotizaciones en bolsa o variables médicas como el nivel de estrés en pacientes. En otras áreas de aplicación, el argumento de las funciones observadas es una magnitud distinta al tiempo tal como pueden ser la localización espacial, la longitud de onda o la probabilidad. En quimiometría existen aplicaciones donde se disponen de observaciones del espectro NIR en una malla fina longitudes de onda.

El análisis de datos funcionales (ADF) es un tema de actualidad en Estadística que cuenta con una gran actividad investigadora y que pretende resolver problemas relacionados con la modelización y la predicción estadística de datos funcionales. Una revisión completa sobre los principales métodos del ADF, aspectos computacionales relacionados con su aplicación práctica e importantes ejemplos con datos reales pueden verse en los libros pioneros de Ramsay and Silverman (2005, 2002); Ramsay et al. (2009). En Ferraty and Vieu (2006) se puede ver un estudio detallado sobre metodologías no paramétricas en ADF. Recientemente, en Horvath and Kokoszka (2012) se ha llevado a cabo inferencia estadística sobre distintos métodos del ADF.

Los primeros trabajos sobre ADF se desarrollaron en el contexto de los

procesos estocásticos en tiempo continuo, generalizando al caso funcional técnicas de reducción de la dimensión tales como el análisis en componentes principales (PCA) (Deville, 1974). Más tarde, las investigaciones estadísticas sobre ADF se centraron en la formulación y estimación de distintos modelos de regresión funcional. Uno de los primeros modelos de regresión extendidos al ámbito de los datos funcionales fue el modelo lineal funcional, que permite estimar una variable escalar a partir de un predictor funcional (Cardot et al., 1999, 2003). En Chiou et al. (2004) se estudió el caso concreto en que el predictor es un vector y la respuesta es funcional. Con objeto de modelizar la media de una variable respuesta funcional en términos de una variable categórica se introdujo el análisis de la varianza funcional (Cuevas et al., 2002, 2004). Por otro lado, en Yao et al. (2005b) y Ocaña et al. (2008) se estudió el modelo lineal funcional para el caso en que tanto la variable respuesta como las variables explicativas son funcionales. Los modelos de regresión en componentes principales, los cuales pueden ser vistos como un caso particular de los anteriores, se utilizaron por primera vez en Aguilera et al. (1997, 1999) para predecir un proceso estocástico en tiempo continuo sobre un intervalo futuro a partir de su pasado más reciente. Por otro lado, los modelos lineales generalizados también se extendieron para el caso de un predictor funcional (James, 2002; Müller, 2005). Un caso particular del modelo lineal generalizado funcional es el modelo de regresión logística funcional, cuyo objetivo es predecir una variable aleatoria binaria a partir de un predictor funcional (Ratcliffe et al., 2002; Escabias et al., 2004; Aguilera et al., 2008b).

La estimación directa de la función parámetro asociada a un modelo de regresión funcional es un problema de difícil solución debido a la dimensión infinita de la variable funcional. Por otro lado, las curvas muestrales usualmente se observan en un conjunto finito de puntos muestrales que pueden ser desigualmente espaciados y diferentes para las distintas unidades muestrales. Por ello, el primer paso en ADF es reconstruir la verdadera forma funcional de cada curva muestral a partir de un conjunto finito de observaciones discretas. Para resolver este problema se han utilizado distintas técnicas de aproximación, tales como la interpolación o la proyección en un espacio finito-dimensional generado por funciones básicas. De este modo, la estimación de un modelo de regresión funcional se reduce a la estimación de un modelo de regresión multivariante equivalente con gran correlación entre las variables predictoras.

Para obtener una estimación adecuada de los parámetros asociados a un

modelo de regresión, usualmente se lleva a cabo una regresión sobre un conjunto de variables aleatorias incorreladas. El ACP funcional se ha utilizado en muchas ocasiones para reducir la dimensionalidad de un conjunto de datos y resolver el problema de multicolinealidad de diversos modelos de regresión funcional. Las componentes principales son combinaciones lineales generalizadas incorreladas de un predictor funcional con varianza máxima. Por ello, la principal crítica sobre la regresión en componentes principales es que los regresores se obtienen sin tener en cuenta la variable respuesta. Con objeto de resolver este problema, el criterio de mínimos cuadrados parciales (PLS) fue extendido al caso funcional para obtener un conjunto de combinaciones lineales generalizadas incorreladas de la variable predictora que tengan covarianza máxima con la variable respuesta (Preda and Saporta, 2005b).

En muchas aplicaciones los datos son funciones suaves observadas con error. En este caso, los coeficientes de la representación básica de la trayectorias muestrales se aproximan mediante mínimos cuadrados con bases de B-splines. El problema es que las curvas muestrales aproximadas (splines de regresión) no controlan el grado de suavidad. Como consecuencia, las componentes principales estimadas y los parámetros funcionales asociados con los modelos de regresión funcional son difíciles de interpretar porque tienen mucha variabilidad y falta de suavidad.

El objetivo general de esta tesis es mejorar la estimación de las metodologías del ADF para el caso de datos funcionales observados con error. Con objeto de resolver este problema, se proponen distintas aproximaciones basadas en la estimación penalizada mediante representación básica de las curvas muestrales con bases de B-splines. Este objetivo general se consigue mediante cinco objetivos específicos:

1. Revisión y comparación de los métodos existentes para la aproximación de curvas suaves con bases de B-splines.
2. Mejorar la estimación del PCA funcional introduciendo distintas aproximaciones basadas en la penalización spline.
3. Desarrollar distintas aproximaciones penalizadas para estimar el modelo logit funcional, utilizando una estimación spline penalizada del PCA funcional.
4. Proponer distintas estimaciones penalizadas para la regresión PLS funcional.

5. Desarrollar una aplicación de las metodologías de estimación penalizada propuestas para mejorar la calidad en la industria alimentaria.

De acuerdo con los objetivos específicos, la tesis se divide en cinco capítulos, cada uno con su correspondiente metodología y resultados. Los contenidos de cada capítulo se han incluido en distintos artículos de investigación actualmente sometidos o aceptados para su publicación en revistas del JCR. Además de las contribuciones metodológicas de cada capítulo, los distintos métodos de estimación penalizada propuestos se han aplicado sobre distintos conjuntos de datos reales y simulados, desarrollando el código de programación necesario y utilizando el software libre R (<http://www.r-project.org>). En el Anexo A al final de esta memoria se ha incluido un resumen detallado de las principales librerías y funciones de R utilizadas en el desarrollo computacional de esta tesis.

E.1 Capítulo 1

La principal propuesta de este capítulo es revisar y comparar tres formas distintas de aproximación de curvas muestrales suaves observadas con error en términos de bases de B-splines: splines de regresión (aproximación mediante mínimos cuadrados no penalizados), splines de suavizado (penalización continua basada en la integral del cuadrado de las derivadas de orden d de cada curva muestral) y P-splines (penalización discreta basada en las diferencias de orden d entre los coeficientes adyacentes de los B-splines). El comportamiento de estas tres aproximaciones spline suavizadas se han estudiado y comparado mediante un estudio de simulación y varias aplicaciones con datos reales. Se han adaptado los métodos de validación cruzada leave-one-out (CV) y validación cruzada generalizada (GCV) con objeto de seleccionar un parámetro de suavizado común para todas las curvas muestrales aproximadas mediante penalización de la falta de suavidad.

La aproximación con bases de B-splines de funciones suaves observadas con ruido se ha utilizado en una amplia variedad de metodologías del ADF. Esto justifica la importancia de comparar las principales aproximaciones suaves en términos de B-splines y extraer conclusiones que permitan a los profesionales usar la herramienta más adecuada en cada caso.

Sea $\{x_i(t) : t \in T, i = 1, \dots, n\}$ una muestra de tamaño n de una variable funcional X , cuyas observaciones son realizaciones independientes e

igualmente distribuidas de un proceso estocástico de segundo orden $X = \{X(t) : t \in T\}$, continuo en media cuadrática, cuyas funciones muestrales pertenecen al espacio de Hilbert $L^2(T)$ de funciones de cuadrado integrable, con el producto escalar usual dado por

$$\langle f, g \rangle = \int_T f(t) g(t) dt, \quad \forall f, g \in L^2(T).$$

Se asume que las trayectorias muestrales pertenecen a un espacio finito-dimensional generado por una base $\{\phi_1(t), \dots, \phi_p(t)\}$ de modo que se expresan como

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t) = a_i' \phi_j(t), \quad i = 1, \dots, n,$$

donde $a_i = (a_{i1}, \dots, a_{ip})'$ es el vector de coeficientes básicos de la i -ésima trayectoria muestral.

En este capítulo, los vectores de coeficientes básicos de las trayectorias muestrales se estiman por distintas metodologías, tales como splines de regresión (sin penalización), splines de suavizado (penalización continua) y P-splines (penalización discreta).

- Splines de regresión: $\hat{a}_i = (\Phi_i' \Phi_i)^{-1} \Phi_i' x_i$, con x_i los vectores de observaciones de las trayectorias muestrales en los nodos de observación y $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$.
- Splines de suavizado: $\hat{a}_i = (\Phi_i' \Phi_i + \lambda R_d)^{-1} \Phi_i' x_i$, donde λ es el parámetro de suavizado y R_d es una matriz cuyos elementos son las integrales de los productos de las derivadas de orden d entre funciones básicas de B-splines.
- P-splines: $\hat{a}_i = (\Phi_i' \Phi_i + \lambda P_d)^{-1} \Phi_i' x_i$, donde $P_d = (\Delta^d)' \Delta^d$, con Δ^d una matriz cuyos elementos son las diferencias de orden d entre coeficientes adyacentes de B-splines.

E.2 Capítulo 2

El análisis de componentes principales funcional (FPCA) es una técnica de reducción de la dimensión que explica la estructura de dependencia de

un conjunto de datos funcionales en términos de variables incorreladas. En muchas aplicaciones los datos son un conjunto de funciones suaves observadas con error. En estos casos, las componentes principales son difíciles de interpretar porque las funciones peso estimadas tienen mucha variabilidad y falta de suavidad. La forma más común de resolver este problema se basa en penalizar la rugosidad de una función a partir de la integral del cuadrado de su derivada de orden d .

En este capítulo, se proponen dos formas alternativas del FPCA penalizado basadas en la expansión básica con B-splines de las curvas muestrales y una penalización P-spline. La principal diferencia entre ambas versiones del FPCA penalizado es que la primera usa la penalización P-spline en el criterio de mínimos cuadrados de estimación de las curvas muestrales en términos de bases de B-splines, mientras que la segunda introduce la penalización P-spline en la restricción de ortonormalidad del algoritmo que calcula las componentes principales. El parámetro de suavizado para ambas versiones del FPCA penalizado se ha obtenido a partir de una adaptación del criterio de validación cruzada (leave-one-out). Con objeto de testar el buen funcionamiento de las aproximaciones penalizadas propuestas y compararlas con el FPCA no penalizado y el *regularized* FPCA, se han llevado a cabo un estudio de simulación y una aplicación con datos funcionales quimiométricos.

En las mismas condiciones del Capítulo 1, sea $\{x_i(t) : t \in T, i = 1, \dots, n\}$ una muestra de funciones correspondientes a la información muestral de una variable funcional X . En general, la j -ésima componente principal se puede expresar en términos de las funciones peso como sigue

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

donde las funciones peso asociadas son las autofunciones del operador de covarianza muestral. Esto es, las soluciones al problema

$$\mathcal{C}(f_j)(t) = \int_T C(t, s) f_j(s) ds = \lambda_j f_j(t).$$

Considerando la representación básica de las curvas muestrales, las funciones peso se pueden expresar en términos de la misma base tal que

$$f_j(t) = \sum_{k=1}^p b_{jk} \phi_k(t) = \phi(t)' b_j,$$

con $b_j = (b_{j1}, \dots, b_{jp})'$. La estimación del vector de coeficientes básicos de las funciones peso depende del tipo de ACPF usado.

- FPCA no penalizado es equivalente a un PCA multivariante de la matriz $A\Psi^{\frac{1}{2}}$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión y $\Psi^{\frac{1}{2}}$ la raíz cuadrada de la matriz de productos interiores entre funciones básicas (Ocaña et al., 2007).
- FPCA de los P-splines es equivalente a un PCA multivariante de la matriz $A\Psi^{\frac{1}{2}}$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con P-splines.

Con objeto de obtener el valor óptimo de λ , en este capítulo se han considerado y comparado dos criterios de selección distintos: validación cruzada leave-one-out (CV) y validación cruzada generalizada (GCV).

- FPCA con suavizado P-spline se reduce a un PCA multivariante de la matriz $A\Psi L^{-1}$, donde A es la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión, L^{-1} es la inversa de una matriz triangular inferior obtenida con la descomposición de Choleski de $LL' = \Psi + \lambda P_d$, con λ el parámetro de suavizado y P_d la matriz de penalización discreta definida en el Capítulo 1.

Para seleccionar el parámetro de suavizado en esta versión del FPCA penalizado, se ha adaptado el criterio de validación cruzada leave-one-out (CV) considerando las distancias cuadráticas en términos de representaciones básicas. Esto consiste en seleccionar el valor de λ que minimiza

$$CV(\lambda) = \frac{1}{p} \sum_{q=1}^p CV_q(\lambda),$$

donde

$$CV_q(\lambda) = \frac{1}{n} \sum_{i=1}^n \|x_i - x_i^{q(-i)}\|^2,$$

con $x_i^{q(-i)} = \sum_{l=1}^q \xi_{il}^{(-i)} f_l^{(-i)}$ la reconstrucción de la curva muestral x_i en términos de las primeras q componentes principales estimadas a partir de una muestra de tamaño $n - 1$ que contiene todas las curvas muestrales excepto x_i .

En términos de expansiones básicas, estas distancias cuadráticas vienen dadas por

$$\|x_i - x_i^{q(-i)}\|^2 = \int_T [x_i(t) - x_i^{q(-i)}(t)]^2 dt = d_i' \Psi d_i,$$

donde $d_i = (d_{i1}, \dots, d_{ip})'$, con $d_{ij} = a_{ij} - \sum_{l=1}^q \xi_{il}^{(-i)} b_{lj}^{(-i)}$, y Ψ la matriz de productos interiores entre las bases de B-splines.

E.3 Capítulo 3

En este capítulo el objetivo es mejorar la estimación del modelo logit funcional. El problema de multicolinealidad asociado con la estimación de este modelo se puede solucionar usando como variables predictoras un conjunto de componentes principales. El parámetro funcional estimado por la regresión logit en componentes principales a veces es ruidosa y difícil de interpretar. Para resolver estos problemas, en este capítulo se proponen distintas estimaciones spline penalizadas del modelo logit funcional. Todas ellas se basan en el FPCA penalizado y/o una penalización P-spline discreta en el criterio de máxima verosimilitud en términos de la expansión básica con B-splines de la curvas muestrales y la función parámetro.

En el contexto de la regresión logística funcional en componentes principales, se introducen tres versiones distintas de estimación penalizada basadas en el FPCA suavizado. Por un lado, se lleva a cabo un FPCA sobre la aproximación P-spline de las curvas muestrales (Método II). Por otro lado, se incluye una penalización P-spline discreta, que penaliza la rugosidad de las funciones peso asociadas a las componentes principales, en la propia formulación del FPCA (Método III). La tercera aproximación penalizada se lleva a cabo introduciendo la penalización en la estimación por máxima verosimilitud de la función parámetro y en términos de un conjunto reducido de componentes principales funcionales (Método IV). Además, se propone un estimación directa (sin componentes principales) de la función parámetro por máxima verosimilitud penalizada en términos de funciones B-spline (Método V). La capacidad de estas aproximaciones para proporcionar una buena estimación de la función parámetro, así como su capacidad de clasificación se comprueba y se compara con la aproximación basada en el ACPF no penalizado mediante un estudio amplio de simulación. Para seleccionar el parámetro de suavizado y el número de componentes principales o funciones básicas, se han adaptado

los criterios de validación cruzada leave-one-out (CV) y validación cruzada generalizada (GCV).

En las mismas condiciones del Capítulo 1, sea $\{x_i(t) : t \in T, i = 1, \dots, n\}$ una muestra de funciones que corresponden a la información muestral de una variable funcional X e $\{y_i : i = 1, \dots, n\}$ una muestra aleatoria de Y asociada a ellas. Esto es, $y_i \in \{0, 1\}, \forall i = 1, \dots, n$. El modelo de regresión logit funcional viene dado por

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

con las transformaciones logit expresadas como

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

Se consideran las representaciones básicas de las trayectorias muestrales y la función parámetro

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad \beta(t) = \sum_{k=1}^p \beta_k \phi_k(t),$$

con $\beta = (\beta_1, \dots, \beta_p)'$ el vector de coeficientes básicos de $\beta(t)$. En este capítulo se desarrollan distintas formas de estimar el vector de coeficientes básicos β .

1. FLoM en términos de componentes principales (FPCLoM)

En general, el FLoM se puede reescribir en términos de componentes principales funcionales como sigue

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

donde $\Gamma = (\xi_{ij})_{n \times p}$ es la matriz de componentes principales funcionales, γ es el vector de coeficientes del modelo y α es la constante.

Una estimación adecuada del parámetro funcional se obtiene considerando sólo un conjunto q de componentes principales como variables predictoras, de modo que $\Gamma = (\xi_{ij})_{n \times q}$ ($q < p$). Por lo tanto, el vector β de coeficientes básicos viene dado por $\beta = F\gamma$, donde la forma en que se estima F depende del tipo de FPCA utilizado para estimar el modelo funcional y del tipo de máxima verosimilitud empleada (penalizada o no penalizada).

- Método I: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, donde G es la matriz cuyas columnas son los autovectores de la matriz de covarianzas muestral de $A\Psi^{1/2}$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión. En este método γ se estima por máxima verosimilitud no penalizada.
- Método II: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, donde G es la matriz cuyas columnas son los autovectores de la matriz de covarianzas muestral de $A\Psi^{1/2}$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con P-splines. En este método γ se estima por máxima verosimilitud no penalizada.
- Método III: $\hat{\beta} = (L^{-1})' G \hat{\gamma}$, donde G es la matriz cuyas columnas son los autovectores de la matriz de covarianzas muestral de $A\Psi(L^{-1})'$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión y L definida en el Capítulo 2. En este método γ se estima por máxima verosimilitud no penalizada.
- Método IV: $\hat{\beta} = \Psi_{p \times p}^{-\frac{1}{2}} G_{p \times n} \hat{\gamma}$, donde G es la matriz cuyas columnas son los autovectores de la matriz de covarianzas muestral de $A\Psi^{1/2}$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión. En este método γ se estima por máxima verosimilitud penalizada

$$\mathcal{L}^*(\lambda, \gamma) = \mathcal{L}(\gamma) - \frac{\lambda}{2} \gamma' P_d \gamma,$$

con $\mathcal{L}(\gamma)$ la log-verosimilitud del FPCLoM, λ el parámetro de suavizado y P_d la matriz de penalización discreta definida en el Capítulo 1.

2. FLoM via log-verosimilitud penalizada.

Se considera la representación básica con B-splines de las trayectorias muestrales y del parámetro funcional. De este modo, las transformaciones logit en forma matricial vienen dadas por

$$L = X\beta,$$

donde $L = (l_1, \dots, l_n)$ es el vector de transformaciones logit, $X = (\mathbf{1}|A\Psi)$, con $\mathbf{1} = (1, \dots, 1)'$ un vector n -dimensional de unos y $\beta = (\beta_1, \dots, \beta_p)'$ el vector de coeficientes básicos de $\beta(t)$.

Método V: el parámetro funcional se estima mediante máxima verosimilitud penalizada y sin componentes principales. Así, la log-verosimilitud penalizada viene dada por

$$\mathcal{L}^*(\lambda, \beta) = \mathcal{L}(\beta) - \frac{\lambda}{2} \beta' P_d \beta,$$

con $\mathcal{L}(\gamma)$ la log-verosimilitud del FLoM. En este caso, la solución de Newton-Raphson para los estimadores de verosimilitud penalizada sería

$$\beta^{(t)} = \beta^{(t-1)} + [X' \text{Diag}(\pi_i^{(t-1)} (1 - \pi_i^{(t-1)})) X + \lambda P_d]^{-1} X' (y - \pi_i^{(t-1)}),$$

con $X = (\mathbf{1} | A\Psi)$.

E.4 Capítulo 4

Los principales problemas asociados con el modelo logit funcional para una respuesta escalar en términos de curvas suaves observadas con error son la alta dimensionalidad, la multicolinealidad y la falta de suavidad en la estimación del parámetro funcional. Con objeto de resolver los tres problemas al mismo tiempo, se proponen dos aproximaciones penalizadas distintas basadas en la regresión PLS. La primera introduce la penalización en la definición de la norma de las funciones peso asociadas a las componentes PLS (Método II). La segunda considera una estimación penalizada de la covarianza entre la respuesta y las componentes PLS (Método III). Se considera tanto penalización continua como discreta basadas ambas en la expansión básica con B-spline de las curvas muestrales. La selección del número óptimo de componentes PLS y del parámetro de suavizado se lleva a cabo mediante dos criterios diferentes basados en los errores de GCV y los errores cuadráticos medios integrados respecto de la función parámetro.

El buen funcionamiento de las aproximaciones penalizadas del FPLS se ha comprobado y comparado con el FPLS no penalizado mediante un estudio de simulación y una aplicación con datos funcionales quimiométricos.

En las mismas condiciones del Capítulo 1, sea $\{x_i(t) : t \in T, i = 1, \dots, n\}$ una muestra de funciones relativas a la información muestral de una variable funcional X y $\{y_i : i = 1, \dots, n\}$ una muestra aleatoria de Y asociada a ellas. El modelo lineal funcional (FLM) viene dado por

$$y_i = \beta_0 + \int_T x_i(t) \beta(t) dt + \varepsilon_i,$$

con $\{\varepsilon_i : i = 1, \dots, n\}$ errores aleatorios centrados e independientes.

Con objeto de conseguir una estimación adecuada del parámetro funcional, el FLM se considera en términos de componentes PLS tal que

$$Y = \mathbf{1}\gamma_0 + T\gamma,$$

donde T es la matriz cuyas columnas son las componentes PLS y γ es el vector de coeficientes de regresión de Y sobre T .

Una estimación adecuada del parámetro funcional se obtiene considerando sólo un conjunto óptimo de q componentes PLS como variables explicativas, de modo que $T = (t_{ij})_{n \times q}$ ($q < p$). La matriz de componentes PLS T se estima con tres versiones distintas de PLS funcional.

Se considera la representación básica de las trayectorias muestrales y de la función parámetro

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad \beta(t) = \sum_{k=1}^p \beta_k \phi_k(t).$$

Así, se consideran las siguientes versiones de regresión PLS funcional (FPLS):

- FPLS no penalizado.

La regresión PLS funcional de una variable respuesta aleatoria real Y en términos de un predictor funcional $X = \{X(t) : t \in T\}$ es un procedimiento iterativo de modo que la h -ésima componente PLS viene dada por $t_h = \int_T X_{h-1}(t) w_h(t) dt$, con $w_h(t)$ la función peso obtenida resolviendo el siguiente problema:

$$\begin{aligned} \max_w \quad & \text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right), \\ & \|w\|^2 = 1 \end{aligned}$$

donde $X_0(t) = X(t)$, $\forall t \in T$ e $Y_0 = Y$. El h -ésimo paso del PLS finaliza con la regresión lineal de $X_{h-1}(t)$ e Y_{h-1} sobre t_h tal que

$$\begin{aligned} X_h(t) &= X_{h-1}(t) - p_h(t) t_h, \quad t \in T \\ Y_h &= Y_{h-1} - c_h t_h, \end{aligned}$$

donde $p_h(t) = (\mathbb{E}(X_{h-1}(t) t_h) / \mathbb{E}(t_h^2))$ y $c_h = (\mathbb{E}(Y_{h-1} t_h) / \mathbb{E}(t_h^2))$.

El FPLS no penalizado se reduce a un PLS multivariante de Y sobre la matriz $A\Psi^{1/2}$ de modo que $T = A\Psi^{1/2}V$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión y V la matriz cuyas columnas son los autovectores asociados a las componentes PLS estimadas.

El vector de coeficientes básicos de la función parámetro β se estima como $\hat{\beta} = (\Psi^{-1/2})'V\hat{\gamma}$.

- FPLS penalizando la norma.

En este caso, la h -ésima componentes PLS penalizada se obtiene mediante el siguiente problema de maximización

$$\max_w \frac{\text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right)}{\langle w, w \rangle + \lambda \text{PEN}_d(w)},$$

donde $\langle w, w \rangle$ es el producto escalar clásico, $\text{PEN}_d(w)$ es una penalización general definida en el Capítulo 1 y λ el parámetro de suavizado.

Esta versión del FPLS penalizado se reduce a un PLS multivariante de Y sobre la matriz $A\Psi(L^{-1})'$ tal que $T = A\Psi(L^{-1})'V$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión, V la matriz cuyas columnas son los autovectores asociados a las componentes PLS estimadas y L una matriz definida en el Capítulo 2.

El vector de coeficientes básicos de la función parámetro β se estima como $\hat{\beta} = (L^{-1})'V\hat{\gamma}$.

- FPLS penalizando la covarianza.

En este caso, la h -ésima componente PLS penalizada se obtiene a partir del siguiente problema de maximización

$$\max_w \frac{\text{Cov}^2 \left(\int_0^T X_{h-1}(t) w(t) dt, Y_{h-1} \right) - \lambda \text{PEN}_d(w)}{\langle w, w \rangle},$$

donde $\langle w, w \rangle$ es el producto escalar clásico, $\text{PEN}_d(w)$ es una penalización general definida en el Capítulo 1 y λ el parámetro de suavizado.

En esta versión del FPLS penalizado la matriz de componentes PLS viene dada por $T = A\Psi(\Psi^{-1/2})'V$, con A la matriz de los coeficientes básicos de las trayectorias muestrales aproximadas con splines de regresión y V la

matriz cuyas columnas son los autovectores asociados a las componentes PLS estimadas. En este caso, el vector de coeficientes básicos β del parámetro funcional se estima como $\hat{\beta} = (\Psi^{-1/2})'V\hat{\gamma}$.

E.5 Capítulo 5

El objetivo de este capítulo es mejorar la calidad en la producción de galletas clasificándolas como buenas o malas a partir de las curvas de resistencia de la masa observada durante el proceso de horneado. Como la variable predictora es funcional, se proponen métodos de clasificación funcional, tales como la regresión logit funcional y el análisis discriminante lineal funcional. Con objeto de mejorar la capacidad de clasificación de estos modelos y obtener una buena estimación de la relación entre la calidad de las galletas y la resistencia de la masa, se propone una aproximación P-spline de las curvas muestrales. Una vez estimado el modelo logit funcional se hace inferencia sobre la función parámetro y los correspondientes cocientes de ventajas haciendo uso de la normalidad asintótica de los estimadores de máxima verosimilitud bajo las condiciones clásicas de regularidad. Finalmente, los resultados sobre clasificación se comparan con otra metodología alternativa como es la clasificación componente a componente (componentwise classification) sobre el modelo logit.

Se considera el problema de clasificación de una muestra de observaciones funcionales $\{x_i(t) : t \in T; i = 1, \dots, n\}$ de acuerdo a una variable respuesta binaria $Y \in \{0, 1\}$ cuyas observaciones se denotan por $\{y_i : i = 1, \dots, n\}$.

1. FPCLoM penalizado (Método II Capítulo 3)

Recordemos el FLoM definido en el Capítulo 3 y las transformaciones logit expresadas como

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

- Interpretación de la función parámetro

$$\exp(l_i^* - l_i) = \exp \left(K \int_{t_0}^{t_0+h} \beta(t) dt \right),$$

donde l_i^* es la transformación logit para la i -ésima observación funcional incrementada de forma constante en el periodo $[t_0, t_0+h]$, y

K una constante positiva. Esto es un cociente de ventajas tal que la ventaja a favor de que $Y = 1$ se multiplica por una cantidad cuando el valor de la observación funcional se incrementa de forma constante en K unidades en un intervalo fijado $[t_0, t_0+h]$.

- Inferencia sobre la función parámetro

Sea $\hat{I} = \int_{t_0}^{t_0+h} \hat{\beta}(t) dt$ el estimador de máxima verosimilitud para $I = \int_{t_0}^{t_0+h} \beta(t) dt$, con $\hat{\beta}(t)$ el estimador de máxima verosimilitud de $\beta(t)$.

Un intervalo de confianza al nivel $(1-\alpha)$ para el cociente de ventajas vendría dado por

$$\exp(\hat{I} \pm \hat{\sigma}(\hat{I}) z_{\alpha/2}).$$

Un intervalo de confianza al nivel $(1-\alpha)$ para $\beta(t)$ vendría dado por

$$\hat{\beta}(t) \pm \hat{\sigma}(\hat{\beta}(t)) z_{\alpha/2}.$$

Consideremos la representación básica de las trayectorias muestrales y el parámetro funcional. Así, el vector de coeficientes básicos de $\beta(t)$ se estima (Method II Chapter 3) como $\hat{\beta} = F\hat{\gamma}$. Por tanto,

$$Var[\hat{I}] = \int_{t_0}^{t_0+h} \int_{t_0}^{t_0+h} Cov(\hat{\beta}(t), \hat{\beta}(s)) dt ds$$

se puede aproximar por

$$Var[\hat{I}] = \int \int \phi'(t) FCov(\hat{\gamma}) F'\phi(s) dt ds.$$

2. LDA-FPLS penalizado

El análisis discriminante lineal (LDA) en el contexto de los datos funcionales consiste en encontrar combinaciones lineales $\int_T X(t) \beta(t) dt$, con $\beta \in L_2(T)$ de modo que la varianza entre clases se maximiza respecto a la varianza total

$$\max_{\beta} \frac{Var(E[X(t)|Y])}{Var(X(t))}.$$

La función discriminante es la función parámetro de la regresión lineal funcional de Y sobre $\{X(t) : t \in T\}$, donde Y se transforma del

siguiente modo

$$\begin{aligned} Y &= -\sqrt{p_0/p_1} \quad \text{si } Y=1, \\ Y &= \sqrt{p_1/p_0} \quad \text{si } Y=0, \end{aligned}$$

con $p_0 = P[Y = 0]$ y $p_1 = P[Y = 1]$.

Debido a la equivalencia entre el análisis discriminante lineal y la regresión lineal, el problema de la gran dimensionalidad de los datos se resuelve utilizando un conjunto de componentes PLS funcionales como variables predictoras. Se considera que las curvas muestrales están afectadas por cierto ruido. Por ello, la función discriminante estimada no es suave y por lo tanto, se propone una estimación suave de la misma basada en una aproximación previa de las curvas muestrales mediante P-splines.

3. *Componentwise classification* basada en el modelo logit

La clasificación componente a componente consiste en determinar un conjunto de puntos relativamente pequeño $\{t_1^*, \dots, t_p^* \in T\}$, que tenga una influencia importante para la clasificación, y aplicar un método de clasificación estándar sobre el vector $(X(t_1^*), \dots, X(t_p^*))$. Un estudio completo sobre las propiedades teóricas de este método y su comportamiento para distintos clasificadores se llevó a cabo en Delaigle et al. (2012). En este capítulo se muestra esta aproximación para el clasificador basado en el modelo logit. La aproximación resultante tiene dos pasos fundamentales basados en la selección del valor de p y la posición del conjunto óptimo de nodos.

Bibliography

- Adams, R. A. (1975). *Sobolev spaces*. Academic Press.
- Agresti, A. (1990). *Categorical data analysis*. Wiley.
- Aguilera, A. M. and Aguilera-Morillo, M. C. (2012). Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling, under revision*.
- Aguilera, A. M. and Aguilera-Morillo, M. C. (2013). Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation, in press*.
- Aguilera, A. M., Aguilera-Morillo, M. C., Escabias, M., and Valderrama, M. J. (2010a). Different P-spline approaches for smoothed functional principal component analysis. In *Proceedings in Computational Statistics (Lechevalier, Y. and Saporta, G., editors)*. Physica-Verlag, pages 641–648.
- Aguilera, A. M., Aguilera-Morillo, M. C., Escabias, M., and Valderrama, M. J. (2011). Penalized spline approaches for functional principal component logit regression. In *Recent Advances in Functional Data Analysis and Related Topics (F. Ferraty, editor)*. Physica-Verlag, pages 1–7.
- Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010b). Using basis expansion for estimating functional PLS regression. Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104:289–305.
- Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2008a). Discussion of different logistic models with functional data. application to Systemic Lupus Erythematosus. *Computational Statistics and Data Analysis*, 53(1):151–163.

- Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2008b). Forecasting binary longitudinal data by a functional PC-ARIMA model. *Computational Statistics and Data Analysis*, 52(6):3187–3197.
- Aguilera, A. M., Gutiérrez, R., Ocaña, F. A., and Valderrama, M. J. (1995). Computational approaches to estimation in the principal component analysis of a stochastic process. *Applied Stochastic Models and Data Analysis*, 11(4):279–299.
- Aguilera, A. M., Gutiérrez, R., and Valderrama, M. J. (1996). Approximation of estimators in the PCA of a stochastic proces using B-splines. *Communications in Statistics. Simulation and Computation*, 25(3):671–690.
- Aguilera, A. M., Ocaña, F. A., and Valderrama, M. J. (1997). An approximated principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis*, 13:61–72.
- Aguilera, A. M., Ocaña, F. A., and Valderrama, M. J. (1999). Forecasting with unequally spaced data by a functional principal component approach. *Test*, 8(1):233–254.
- Aguilera-Morillo, M. C. and Aguilera, A. M. (2012). P-spline estimation of functional classification methods for improving the quality in the food industry. *Communications in Statistics - Simulation and Computation, under revision*.
- Aguilera-Morillo, M. C., Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2012). Penalized spline approaches for functional logit regression. *TEST, in press*, 51.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.
- Bali, J., Boente, G., Tyler, D., and Wang, J. (2011). Robust functional principal components: a projection-pursuit approach. *Annals of Statistics*, 39:2852–2882.
- Besse, P., Cardot, H., and Ferraty, F. (1997). Simultaneous nonparametric regression of unbalanced longitudinal data. *Computational Statistics and Data Analysis*, 24:255–270.

- Besse, P. and Ramsay, J. O. (1986). Principal component analysis of sample functions. *Psychometrika*, 51(2):285–311.
- Bhatti, M. and Bracken, P. (2006). The calculation of integrals involving B-splines by means of recursion relations. *Applied Mathematics and Computation*, 172:91–100.
- Boente, G. and Fraiman, R. (2000). Kernel-based functional principal components. *Statistics and Probability Letters*, 48:335–345.
- Bosq, D. (2000). *Linear processes in function spaces. Theory and applications. Lecture notes in Statistics*, 149. Springer-Verlag.
- Bosq, D. and Blanke, D. (2007). *Inference and predictions in large dimensions*. Jhon Wiley and Sons. Paris.
- Bouzas, P. R., Aguilera, A. M., Valderrama, M. J., and Ruiz-Fuentes, N. (2006). Modelling the mean of a doubly stochastic poisson process by functional data analysis. *Computational Statistics and Data Analysis*, 50:2655–2667.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.
- Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443):961–976.
- Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics*, 34:2159–2179.
- Camacho, J., Picó, J., and Ferrer, A. (2010). Data understanding with PCA: Structural and variance information plots. *Chemometrics and Intelligent Laboratory Systems*, 100(1):48–56.
- Cardot, H., Crambes, C., Kneip, A., and Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis*, 51:4832–4848.
- Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, 45:11–22.

- Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.
- Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41.
- Castro, P. E., Lawton, W. H., and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28:329–337.
- Chiou, J. M., Müller, H. G., and Wang, J. L. (2004). Functional response models. *Statistica Sinica*, 14:659–677.
- Crambes, C., Kneip, A. F., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of statistics*, 37:35–72.
- Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403.
- Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canadian Journal of Statistics*, 30:285–300.
- Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics and Data Analysis*, 47(2):111–122.
- Currie, I. and Durban, M. (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, 2:333–349.
- Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12:136–156.
- De Boor, C. (1977). Package for calculating with B-splines. *Journal of Numerical Analysis*, 14:441–472.
- De Boor, C. (2001). *A practical guide to splines (revised edition)*. Springer.
- Delaigle, A. and Hall, P. (2012a). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society. Series B*, 74(2):267–286.

- Delaigle, A. and Hall, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1):322–352.
- Delaigle, A., Hall, P., and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, 99(2):299–313.
- Deville, J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15:3–101.
- Durban, M. (2007). *Splines con penalizaciones: Teoría y aplicaciones*. Universidad Pública de Navarra.
- Durban, M. (2009). An introduction to Smoothing with Penalties: P-splines. *Boletín de la sociedad Española de Estadística e Investigación Operativa (BEIO)*, 25(3):195–205.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.
- Eilers, P. H. C. and Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, 11:758–783.
- Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews Computational Statistics*, 2:637–653.
- Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384.
- Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 16(1):95–107.
- Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2007). Functional PLS logit regression model. *Computational Statistics and Data Analysis*, 51(10):4891–4902.
- Escabias, M., Valderrama, M. J., Aguilera, A. M., Satofimia, M. E., and Aguilera-Morillo, M. C. (2013). Stepwise selection of functional covariates in forecasting peak levels of olive pollen. *Stochastic Environmental Research and Risk Assessment*, 27:367–376.

- Escabias, M., Valderrama, M. J., and Aguilera-Morillo, M. C. (2012). Functional Data Analysis in Biometrics and Biostatistics. *Journal of Biometrics and Biostatistics*, 3(8):1–2.
- Ferraty, F., Hall, P., and View, P. (2010). Most-predictive design points for functional data predictors. *Biometrika*, 97:807–824.
- Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, 44:161–173.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis. Theory and practice*. Springer-Verlag.
- Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling (with discussion and response). *Technometrics*, 31:1–39.
- García Nieto, P. J., Martínez Torres, J., de Cos Juez, F. J., and Sánchez Lasheras, F. (2012). Using multivariate adaptive regression splines and multilayer perceptron networks to evaluate paper manufactured using *Eucalyptus globulus*. *Applied Mathematics and Computation*, 219:755–763.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. Monographs on Statistics and applied probability. Chapman & Hall.
- Guo, W. (2004). Functional data analysis in longitudinal settings using smoothing splines. *Statistical Methods in Medical Research*, 13(1):49–62.
- Harezlak, J., Coull, B. A., Laird, N. M., Magari, S., and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational Statistics and Data Analysis*, 51(10):4911–4925.
- Hastie, T., Buja, A., and Tibshirani, R. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society. B*, 55:757–796.

- Horvath, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. Springer-Verlag.
- Huang, J. Z., , Shen, H., and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2:678–695.
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley.
- James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society. Series B*, 64(3):411–432.
- James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):533–550.
- James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87:587–602.
- Kalivas, J. H. (1997). Two data sets of Near-Infrared Spectra. *Chemometrics and Intelligent Laboratory Systems*, 37:255–259.
- Kano, H., Fujioka, H., and Martin, C. F. (2011). Optimal smoothing and interpolating splines with constraints. *Applied Mathematics and Computation*, 218:1831–1844.
- Kano, H., Nakata, H., and Martin, C. F. (2005). Optimal curve fitting and smoothing using normalized uniform B-splines: a tool for studying complex systems. *Applied Mathematics and Computation*, 169:96–128.
- Kayano, M. and Konishi, S. (2009). Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data. *Journal of Statistical Planning and Inference*, 139(7):2388–2398.
- Krämer, N., Boulesteix, A.-L., and Tutz, G. (2008). Penalized partial least squares with applications to b-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94:60–69.
- Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

- Lee, T. C. M. (2000). Regression spline smoothing using the minimum description length principle. *Statistics and Probability Letters*, 48:71–82.
- Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., and Cohen, K. (1999). Robust principal components for functional data (with discussion). *Test*, 8:1–73.
- Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves. a P-spline approach. *Technometrics*, 41(1):1–13.
- Müller, H. G. (2005). Functional modelling and classification of longitudinal data. *Board of the Foundation of the Scandinavian Journal of Statistics*, 32:223–240.
- Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.
- Ocaña, F. A., Aguilera, A. M., and Escabias, M. (2007). Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465.
- Ocaña, F. A., Aguilera, A. M., and Valderrama, M. J. (1999). Functional Principal Components Analysis by Choice of Norm. *Journal of Multivariate Analysis*, 71:262–276.
- Ocaña, F. A., Aguilera, A. M., and Valderrama, M. J. (2008). Estimation of functional regression models for functional responses by wavelet approximations. In *Functional and Operatorial Statistics (S. Dabo-Niang and F. Ferraty, editors)*. Physica-Verlag, pages 15–22.
- O’Sullivan, F. (1986). A stastical perspective on ill-posed inverse problems. *Statistical Science*, 1:505–527.
- Pigoli, D. and Sangalli, L. (2012). Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives. *Computational Statistics and Data Analysis*, 56:1482–1498.
- Preda, C. and Saporta, G. (2005a). Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 49:99–108.
- Preda, C. and Saporta, G. (2005b). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48:149–158.

- Preda, C., Saporta, G., and Lévêder, C. (2007). PLS classification for functional data. *Computational Statistics*, 22:223–235.
- Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer-Verlag.
- Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis*. Springer-Verlag.
- Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. Springer-Verlag.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis (Second Edition)*. Springer-Verlag.
- Ramsay, J. O. and Wang, X. (1995). A functional data analysis of the pinch force of human fingers. *Applied Statistics*, 44:17–30.
- Ratcliffe, S. J., Heller, G. Z., and Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression. *Statistics in Medicine*, 21:1115–1127.
- Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10:177–183.
- Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.
- Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14:631–647.
- Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society B*, 53:233–527.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757.
- Saeyes, W., Darius, P., and Ramon, H. (2004). Potential for on-site analysis of hog manure using a visual and near infrared diode array reflectance spectrometer. *Journal of Near Infrared Spectroscopy*, 12:299–309.

- Saeyns, W., De Ketelaere, B., and Darius, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, 22(5):335–344.
- Sangalli, L., Ramsay, J., and Ramsay, T. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society Ser. B, Statistical Methodology*, 75(4):1–23.
- Saporta, G. (1981). Méthodes exploratoires d'analyse de données temporelles. *Cahiers du B.U.R.O., Université Pierre et Marie Curie*, pages 37–38.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annal Statistics*, 6.
- Segovia-Gonzalez, M. M., Guerrero, F. M., and Herranz, P. (2009). Explaining functional principal component analysis to actuarial science with an example on vehicle insurance. *Insurance: Mathematics and Economics*. Elsevier, 45(2):278–285.
- Silverman, B. W. (1996). Smoothed functional principal component analysis by choice of norm. *Annal Statistics*, 24:1–24.
- Valderrama, M. J., Ocaña, F. A., Aguilera, A. M., and Ocaña-Peinado, F. M. (2010). Forecasting pollen concentration by a two-step functional model. *Biometrics*, 66:578–585.
- Van der Linde, A. (2008). Variational bayesian functional PCA. *Computational Statistics and Data Analysis*, 53(2):517–533.
- Van Trees, H. L. (1968). *Detection, estimation and modulation theory: Part I*. Wiley.
- Viviani, R., Gron, G., and Spitzer, M. (2005). Functional principal component analysis of fMRI data. *Human Brain Mapping*, 24:109–129.
- Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.
- Yang, W., Müller, H. G., and Stadtmüller, U. (2011). Functional singular component analysis. *Journal of the Royal Statistical Society B*, 73(3):303–324.

- Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society B*, 68:3–25.
- Yao, F., Müller, H. G., and Wang, J. L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of American Statistical Association*, 100:577–590.
- Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33(6):2873–2903.
- Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, 96(453):247–259.

List of Figures

1.1 B-splines bases	20
1.2 Regression spline approaches	25
1.3 Simulation Chapter 1. CV vs GCV	31
1.4 Simulation Chapter 1. Approaches with different number of basis knots	32
1.5 Simulation Chapter 1. Different approaches (Regression splines, smoothing splines and P-splines)	33
1.6 Simulation Chapter 1. Mean function and MSE	34
1.7 Application Chapter 1 (pinch data). Original curves and their approximations	36
1.8 Application Chapter 1 (pinch data). Original curves and P-splines	37
1.9 Application Chapter 1 (manure data). Original curves and their approximations	38
1.10 Application Chapter 1 (manure data): Original curves and P-splines	39
2.1 Simulation Chapter 2. Different approaches (Regression splines, smoothing splines and P-splines)	55
2.2 Simulation Chapter 2. First, second and third eigenfunction with different basis knots	56
2.3 Simulation Chapter 2. Sample paths reconstructions with the first three PCs	57
2.4 Simulation Chapter 2. MSE of the first three eigenfunctions	58
2.5 Simulation Chapter 2. MSE of the sample curves reconstructions with the first three eigenfunctions	59
2.6 Simulation Chapter 2. MSE of the sample curves reconstructions with all eigenfunctions	59

2.7	Application Chapter 2 (diesel data). First and second estimated eigenfunctions	61
2.8	Application Chapter 2 (diesel data). Noisy sample paths and its reconstructions with the two first PCs	61
2.9	Application Chapter 2 (diesel data). Original noisy test sample and its reconstructions with the two first PCs	62
3.1	Simulation Chapter 3 (Case I). Simulated sample curves	79
3.2	Simulation Chapter 3 (Case I). Estimated functional parameter for one of the 100 simulations	80
3.3	Simulation Chapter 3 (Case I). Mean of the parameter function and confidence bands	82
3.4	Simulation Chapter 3 (Case I). Area under ROC curve and MSE distribution	83
3.5	Simulation Chapter 3 (Case II). Mean parameter function and confidence bands (Methods I, II, III, IV, and V)	86
3.6	Simulation Chapter 3 (Case II). Distribution of the $IMSE_{\beta}$ for the estimated parameter functions on 200 repetitions	87
3.7	Simulation Chapter 3 (Case II). Mean of the parameter function and confidence bands (FLDA-PLS)	87
3.8	Simulation Chapter 3 (Case II). Area under ROC curve and MSE distribution	88
4.1	Simulation Chapter 4. Sample data set and regression splines	105
4.2	Simulation Chapter 4. Problems of functional linear model	106
4.3	Simulation Chapter 4. Simulated parameter function	107
4.4	Simulation Chapter 4 (Case I and Case II). Simulated parameter function and mean parameter functions	108
4.5	Simulation Chapter 4 (Case I). Mean parameter functions and confidence bands	112
4.6	Simulation Chapter 4 (Case II). Mean of the parameter functions and confidence bands	113
4.7	Simulation Chapter 4 (Case I and Case II). Box plots related to $IMSE_{\beta}$, GCVE, MSPE and number of PLS components	114
4.8	Application Chapter 4 (gasoline data). Mean of the parameter functions and pointwise confidence bands	116
4.9	Application Chapter 4 (gasoline data). Mean of the parameter functions	117

- 5.1 Application Chapter 5 (DANONE data): Curves of resistance of dough recorded at 240 seconds for 50 flours 125
- 5.2 Application Chapter 5 (DANONE data): Original sample curve, regression splines and P-spline approach of a sample curve for good and bad flour 127
- 5.3 Application Chapter 5 (DANONE data): Means of the parameter functions and the discriminant functions 135
- 5.4 Application Chapter 5 (DANONE data): Box-plots for the distribution of the number of predictors selected for the three considered classifiers 137
- 5.5 Application Chapter 5 (DANONE data): Parameter functions estimated by FPCLoR and P-spline smoothed FPCLoR 138
- 5.6 Application Chapter 5 (DANONE data): Parameter functions estimated by non smoothed and smoothed FPCLoR next to 95% pointwise confidence bands 138
- 5.7 Application Chapter 5 (DANONE data): First, second and third eigenfunction 141
- 5.8 Application Chapter 5 (DANONE data): Mean resistance curves and the effect of adding (+) and subtracting (-) a suitable multiple of each principal component curve 141
- 5.9 Application Chapter 5 (DANONE data): Dispersion graph between the first and the second principal component scores 142

List of Tables

2.1	The first 14 solutions b_i of the Equation (2.3).	53
2.2	Application Chapter 2 (diesel data). Variances and percentages of variances explained by the three considered approaches	60
2.3	Computational cost Chapter 2	63
3.1	Simulation Chapter 3 (Case I). Mean and standard deviation (S.D.) for the GCV errors of the estimated models	81
3.2	Simulation Chapter 3 (Case II). Mean and standard deviation of the $IMSE\beta$	85
4.1	Simulation Chapter 4 (Case I and Case II). Sample mean and standard deviation related to the distribution of $IMSE\beta$, GCVE, MSPE and number of PLS components	111
4.2	Application Chapter 4 (gasoline data): Sample mean and standard deviation of the GCVE, MSPE and the number of PLS components distribution	117
5.1	Application Chapter 5 (DANONE data): Mean and sample standard deviation of the areas under ROC curve and the misclassification rates	136
5.2	Application Chapter 5 (DANONE data): Significance test for the coefficients of the model	140