CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESIS DE DOCTORADO

ADVANCES IN FUNCTIONAL REGRESSION AND CLASSIFICATION MODELS

Manuel Oviedo de la Fuente

ESCUELA DE DOCTORADO INTERNACIONAL

PROGRAMA DE DOCTORADO EN ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

SANTIAGO DE COMPOSTELA

2018

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR DE LA TESIS

Advances in functional regression and classification models

D. Manuel Oviedo de la Fuente

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1) *La tesis abarca los resultados de la elaboración de mi trabajo.*
2) *En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.*
3) *La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.*
4) *Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.*

En Santiago de Compostela, 18 de octubre de 2018.

Fdo Manuel Oviedo de la Fuente

AUTORIZACIÓN DEL DIRECTOR/A DE LA TESIS

Advances in functional regression and classification models

D./Dª. Manuel Febrero Bande

INFORMA/N:

Que la presente tesis, se corresponde con el trabajo realizado por D. Manuel Oviedo de la Fuente, bajo mi dirección, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

De acuerdo con el artículo 41 del Reglamento de Estudios de Doctorado, declara también que la presente tesis doctoral es idónea para ser defendida en base a la modalidad de COMPENDIO DE PUBLICACIONES, en los que la participación del doctorando/a fue decisiva para su elaboración.

La utilización de estos artículos en esta memoria, está en conocimiento de los coautores, tanto doctores como no doctores. Además, estos últimos tienen conocimiento de que ninguno de los trabajos aquí reunidos podrá ser presentado en ninguna otra tesis doctoral.

En Santiago,18 de octubre de 2018

Fdo.: Manuel Febrero Bande

Agradecimientos

No puedo empezar de otra forma que dedicar este trabajo a mis padres. Nunca olvidaré estas disyuntivas entre "una azada" o "un libro", ¡aquí tenéis el mejor que he logrado escribir!

A Manolo, por haberme introducido en el análisis de datos funcionales y en el que el serio trabajo invertido a lo largo de estos años ha revertido en muchas satisfacciones. A Pilar Muñoz, por darme la alternativa profesional como estadístico tanto en el mundo académico como profesional. A los couatores de las publicaciones que conforman esta tesis: Àngela Domínguez, Celestino Ordoñez, Javier Roca Pardiñas, Jose Ramón Rodríguez, Pilar Muñoz, Wenceslao González y Manuel Febrero. A mis compañeros de trabajo, departamento, grupo MODESTYA, cafés y a todos los que, de una forma u otra, han contribuido a la culminación de esta tesis.

Por último, pero no por ello menos importante, gracias a Xandra por tu apoyo y amor incondicional, que conjuntamente con Martín sois los motores de mi vida, ¡gracias a vosotros soy mejor persona!

Fundings Financial Support

This work has been supported by Project MTM2008–03010 from Ministerio de Ciencia e Innovación, MTM2013-41383-P from Ministerio de Economía y Competitividad and MTM2016-76969-P from Ministerio de Economía y Competitividad - Agencia Estatal de Investigación.

Contents

Chapter 1

Introduction

Abstract

Functional data analysis (FDA) has become a very active field of research in the last few years because it appears naturally in most scientific fields: econometrics (curves of financial assets), energy (demand or electricity price curves), environment (curves of pollutant levels), medicine (growth curves), chemometrics (spectrometric data), etc. This thesis is a compendium of the following publications: 1) "Statistical computing in functional data analysis: the R package fda.usc" published in the *J STAT SOFTW*, the core advances of this paper was to propose a common framework in R software where to integrate all the tools of FDA (exploratory, regression, classification,...) and where to implement the new proposals. 2) "Predicting seasonal influenza transmission using functional regression models with temporal dependence" published in *PLoS ONE* proposes an extension of GLS model to functional case. 3) "The DD^G–classifier in the functional setting" published in *TEST* extends the DD-classifier using information derived of the depth of functional data. 4) "Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach" published in *CHEMOMETR INTELL LAB SYST* studies the utility of distance correlation as a method to select impact points in functional regression. 5) "Variable selection in Functional Additive Regression Models", published in *Comput Stat* journal, proposes a variable selection algorithm in the case of functional predictors that may be mixed with other types of variables (scalar, multivariate, directional, etc.). The proposed algorithms have been tested under simulation scenarios and with remarkable datasets in several contexts such as: environment (air pollutant, forest fires), energy (wind power generation) or epidemiology (disease evolution). The methodological developments will be accompanied by practical implementation in the form of new software (R package `fda.usc`) and, in as far as possible, as a technological transfer to the public health services and industry.

Contents

1.1 Manuscript distribution

This section summarizes the main results of the thesis. Each chapter contains an original article with its own abstract, sections, appendices and references. The papers from Chapters 2–6 have been published in peer-reviewed journals with high impact factor. All methodological advances have been presented at national and international conferences in the field of functional data analysis (FDA). Short summaries of all the chapters are given below.

Chapter 2. Statistical Computing in Functional Data Analysis

This chapter is a detailed review of basic state-of-the-art techniques and statistical methods that can be applied to a set of functional data. The core advances of this chapter was to propose a common framework in R (in the form of a package) that agglutinated, in a structured and homogeneous way, the proposals of the two main trends in the analysis of functional data (FDA) and main bibliographic references; the work of Ramsay and Silverman (2005) is related to FDA in Hilbert spaces, and the works of Ferraty and Vieu (2006) and Ferraty and Romain (2010) to more general spaces. This section highlights the work of maintain and integrate (following the standards of the R programming language) the most relevant procedures published in FDA literature with its own developments in the R package `fda.usc`. In particular, the first part of the chapter defines and implements the basic concepts on FDA such as the record/register, the basis representation, smoothing techniques, functional depth, functional outliers, among others. The second part of the chapter is about functional regression models. First, it presents the model with scalar response and functional covariable from the parametric approach as well as from the nonlinear regression. Other relevant procedures, such as the semi-linear model and tools such as identifying influential observations have also been implemented in R.

These developments refer to the article titled "**Statistical computing in functional data analysis: the R package** `fda.usc`" published in the Journal of Statistical Software in 2012 for which the associated information (including article and version 1.0 of the `fda.usc` package) is available in the DOI: `http://dx.doi.org/10.18637/jss.v051.i04`, see Febrero-Bande and Oviedo de la Fuente (2012). Finally, the Appendix A collects the extensions and advances made with respect to the published article, as it is: the logistic regression studied in Escabias et al. (2004), the functional generalized linear model (FGLM) proposed by Müller and Stadtmüller (2005), the extension of the functional generalized spectral additive model (FGSAM) proposed by Müller and Yao (2008), the generalized additive model based on kernel estimators (FGKAM) proposed by Febrero-Bande and González-Manteiga (2013) and the regression model with functional response, see Chiou et al. (2004).

Chapter 3. Regression models with dependent data

The aim of this chapter is to extend the ideas of generalized least squares (GLS) methods discussed in Kariya and Kurata (2004) to functional regression models (FRM) with scalar response. The GLS estimators allow us to incorporate a wide list of covariance structures for the error term into regression models. As an example, these models can include serial correlation, spatial dependence, heteroscedasticity or even, random effects. An iterative version of the GLS estimator (called iGLS) that can help to model complicated dependence structures was also proposed.

The article entitled "**Predicting seasonal influenza transmission using functional regression models with temporal dependence**" published in *PLoS ONE* journal in 2018 includes these developments; its associated information (including the article and version 1.4 of the `fda.usc` package) is available in the DOI: `https://doi.org/10.1371/journal.pone.0194250`, see Oviedo de la Fuente et al. (2018). The procedure was applied to a real problem related with the prediction of the influenza rate in Galicia.

Chapter 4. Functional depth classification
Functional classification based on data depth

Chapter 4 is mainly dedicated to supervised classification methods using classifiers based on the information derived from depth measures of functional data. This chapter aims to extend the DD–classifier for multivariate data (Li et al., 2012) in three ways: first, by enabling it to handle more than two groups; second, by applying regular classification methods (such as *k*NN, linear or quadratic classifiers, recursive partitioning,...) to DD–plots, particularly useful because it gives insights based on the diagnostics of these methods; and third, by integrating various sources of information (data depths, multivariate functional data,...) in the classification procedure in a unified way. This chapter also proposes an enhanced revision of several functional data depths and it provides a simulation study and applications to some real datasets.

This chapter makes reference to the article entitled "**The DD^G-classifier in the functional setting**" published in the journal *TEST* in 2017, available in: `https://link.springer.com/article/10.1007%2Fs11749-016-0502-6`; its supplementary material is available in online version in the DOI: `https://doi.org/10.1007/s11749-016-0502-6`, see Cuesta-Albertos et al. (2017).

Chapter 5. Functional regression with points of impact

In this work we study the utility of distance correlation (Székely et al., 2007) as a method to select impact points for functional data predictor and scalar response. This requires neither projection nor transformation of the impact points. Moreover, it is unnecessary to assume an a priori regression model; we simply look for local maxima of the distance correlation function. The article is structured as follows: First, we provide a brief summary of the functional parametric (Ramsay and Silverman, 2005) and non-parametric regression models (Ferraty and Vieu, 2006). Second, we provide a brief explanation of the three methods used to determine optimum wavelengths (most-predictive design points) for functional data predictor: one is based on a non-parametric kernel smoothing (Ferraty *et al.*, 2010), another is a wavelet-based weighted LASSO regression (Zhao et al., 2015), and our proposal (Ordóñez et al., 2018), based on calculating local maxima on a distance correlation function. Third, we apply all the considered methods to simulated and real data.

This chapter makes reference to the article entitled "**Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach**" published in the journal *Chemometrics and Intelligent Laboratory Systems.* 2017, its supplementary material is available in online version, DOI: `https://doi.org/10.1016/j.chemolab.2017.12.001`, see Ordóñez et al. (2018).

Chapter 6. Variable selection in functional regression models

This chapter considers the problem of variable selection in regression models in the case of different types of predictors (scalar, multivariate, functional, directional, etc.). Our proposal begins with a simple null model and sequentially selects a new variable to be incorporated into the model based on the use of distance correlation proposed by Székely et al. (2007). For multivariate data, we have compared our proposal in the same scenarios of Yenigün and Rizzo (2015) and in a mixed scenario with functional and scalar variables. Furthermore, the last numerical results example is related with a classification problem; so the response is binomial. The procedure was applied to a real problem related with the Iberian Energy Market (Price and Demand) where the number of possible covariates is really big.

This chapter makes reference to the article entitled "**Variable selection in Functional Additive Regression Models**", see Febrero-Bande *et al.* (2018), published in Computational Statistics, DOI: `https://doi.org/10.1007/s00180-018-0844-5`. A preliminary version of this paper was published in the chapter of the same name in the book (Febrero-Bande *et al.*, 2017, pp. 113-122).

1.2 Background

An important technological change has been produced in last years that consists in faster and more accurate equipment that provides more reliable and faster measurements. This technological evolution changes or replaces some of the paradigms on which classical statistics are founded, for example, those in which the number of observations is greater than the number of variables, given a set of data. At the same time, many fields have begun to work with large databases in which it is increasingly common to record the observations of a random variable in a continuous interval. For example, in fields like spectroscopy, the result of the measurement is a curve that has been evaluated in at least one hundred points. This type of data, usually called functional data, naturally arises in many disciplines for instance, we could talk about intra-day stock price curves in energy, leaf wine reflectance curves in agriculture, electricity demand curves in the environment, the evolution of seasonal influenza rate in epidemiology, spectrometric curves in chemistry, etc.

This thesis reviews as its starting point the basic contributions in the literature on functional data corresponding to the books of Ramsay and Silverman (2005) and Ferraty and Vieu (2006) respectively.

The novelty of the functional data analysis (FDA) opens a wide range of potential research lines on the complexity of the data. The recent reviews of Cuevas (2014) and Ferraty et al. (2011) provide a wide perspective on FDA.

The following is a brief introduction on the common concepts for the analysis of functional data developed in this document.

1.2.1 What can be considered a functional data?

FDA deals with the analysis of data in the form of functions such as curves, surfaces, images and shapes, or more general objects. In its simplest form, FDA focuses on the observations of a random variable recorded in a continuous interval (or a finite number of increasingly large discretization points of the continuous interval). The atom of the functional data is a function, where one or more functions are recorded in a random sample. This thesis focuses on taking functions in one of its most common forms in literature, such as curves.

Following the definitions in Ferraty and Vieu (2006):

- **Definition 1.1**. A random variable $\mathcal{X}$ is called a functional variable if it takes values in a functional space $\mathscr{E}$-complete metric (or semi–metric) space-.
- **Definition 1.2**. A functional dataset $\{\mathcal{X}_1,\dots,\mathcal{X}_n\}$ is the observation of n functional variables $\mathcal{X}_1,\dots,\mathcal{X}_n$ identically distributed as $\mathcal{X}$.

Below, we define the basic properties of metric (and semimetric), norm (and seminorm) and the inner product between functional elements. In particular, the functional spaces can be divided into three main categories (in order of complexity).

i. **Metric spaces**, where only the definition of distance is provided:

   A metric on a functional space $\mathcal{F}$ is a map $d(\cdot,\cdot):\mathcal{F}\times\mathcal{F}$ that verifies the following properties for all $\mathcal{X},\mathcal{Y},\mathcal{Z}\in\mathcal{F}$:

   (a) Triangle inequality $d(\mathcal{X},\mathcal{Y})\leq d(\mathcal{X},\mathcal{Y})+d(\mathcal{Y},\mathcal{Z})$

   (b) Symmetry: $d(\mathcal{X},\mathcal{Y})=d(\mathcal{Y},\mathcal{X})$

   (c) Positive definiteness: $d(\mathcal{X},\mathcal{Y})\geq 0$

   (d) Non–degeneracy: $d(\mathcal{X},\mathcal{Y})=0\Longleftrightarrow\mathcal{X}=\mathcal{Y}$

   If $d(\mathcal{X},\mathcal{Y})=0$ does not preclude that $\mathcal{X}\neq\mathcal{Y}$, then $d(\cdot,\cdot)$ is called semi–metric.

ii. **Normed (Banach) spaces**, where also we can define norms:

   A norm in a vectorial space $\mathcal{F}$ is a map $\|\cdot\|:\mathcal{F}\rightarrow\mathbb{R}$ that verifies the following properties for all $\mathcal{X},\mathcal{Y}\in\mathcal{F}$, $\alpha\in\mathbb{R}$:

   (a) Triangle inequality $\|\mathcal{X}+\mathcal{Y}\|\leq\|\mathcal{X}\|+\|\mathcal{Y}\|$

   (b) $\|\alpha\mathcal{X}\|=|\alpha|\|\mathcal{X}\|$

   (c) Si $\|\mathcal{X}\|=0\Rightarrow\mathcal{X}=0$

iii. **Inner product (Hilbert) spaces**, where we can define also an inner product operator as a map $\langle\cdot,\cdot\rangle:\mathcal{F}\times\mathcal{F}\rightarrow\mathbb{R}$ that verifies the following properties for all $\mathcal{X},\mathcal{Y}\in\mathcal{F}$, $\alpha,\beta\in\mathbb{R}$:

   (a) Positive definiteness: $\langle\mathcal{X},\mathcal{X}\rangle\geq 0$ and $\langle\mathcal{X},\mathcal{X}\rangle=0$ for $\mathcal{X}=0$

(b) Symmetry: $\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \mathcal{Y}, \mathcal{X} \rangle$

(c) Linearity: $\langle \alpha \mathcal{X} + \beta \mathcal{X}, \mathcal{Z} \rangle = \alpha \langle \mathcal{X} + \mathcal{X}, \mathcal{Z} \rangle + \beta \langle \mathcal{X} + \mathcal{X}, \mathcal{Z} \rangle$

A Hilbert space $\mathcal{H}$ is always a normed space with the canonical definition of norm from the inner product $\|\mathcal{X}\| = \langle \mathcal{X}, \mathcal{X} \rangle^{1/2}, \forall \mathcal{H}$. An example of Hilbert space is the space $\mathcal{L}^2([a=0, b=1])$ of square–integrable functions with respect to the Lebesgue measure on the unit interval, where the inner product of $\mathcal{X}, \mathcal{Y} \in \mathcal{L}^2([0,1])$ is defined by $\langle \mathcal{X}, \mathcal{Y} \rangle = \int_0^1 X(t)Y(t)dt$ in the real case. While Ramsay and Silverman (2005) are limited to considering the general framework of a Hilbert space, where usually a random variable $\mathcal{X}$ is observed in a discretization points $\{t_j\}_{j=1}^{m} \in T = [a,b] \subset \mathbb{R}$. Then, the functional data can be denoted as $\mathcal{X} = \{X(t); t \in T\}$. Ferraty and Vieu (2006) widens the scope of their analysis to non Hilbertian spaces. All measurements are approximate using numerical approximations that are usually solved by integral and differential calculus. For example, the inner product of $\mathcal{X}, \mathcal{Y}$ can be approximated by trapezoidal rule: $\langle \mathcal{X}, \mathcal{Y} \rangle = \int_a^b X(t)Y(t)dt \approx \frac{1}{2}\sum_{t=1}^{T-1}(t_{i+1}-t_i)\left(X(t_{i+1})Y(t_{i+1}) + X(t_i)Y(t_i)\right)$.

1.3 Exploratory analysis

1.3.1 Representation of functional data

Ramsay and Silverman (2005, p. 43 and 44) considered the approximation of $\mathcal{X}$ by basis functions. A basis function system is as a set of known functions $\{\psi_k\}_{k \in \mathbb{N}}$ such that any function can be approximated arbitrarily well by taking a linear combination of a sufficiently large number of the k_n functions:

$$X(t) = \sum_{k \in \mathbb{N}} c_k \psi_k(t) \approx \sum_{k=1}^{k_n} c_k \psi_k(t) \tag{1.1}$$

The two basis types most commonly used are:

- Fixed basis (functions defined on a grid of knots):
  - Fourier (Ramsay and Silverman, 2005): It is made of trigonometric functions and so, it is recommended to represent periodic functions.
  - B-spline (Cardot et al., 2003): Its design, from a set of polynomials defined in subintervals, makes it flexible and adaptable to problems with smooth curves. It is not orthogonal.
  - Wavelet (Antoniadis and Sapatinas, 2003): It is orthonormal and is widely used in reduction of the dimension in signals due to its design based on translation and dilatation.
- Data–driven basis (functions computed from the data):
  - Functional principal components (FPC), see Cardot et al. (1999). As in multivariate framework, the basis of FPC is obtained sequentially as the projections that maximizes the variance and are orthogonal to the previously obtained ones.
  - Functional partial least squares (FPLS), see Preda and Saporta (2005). It uses the same mechanism as FPC but maximizing the covariance between the response and the predictors.

1.3.2 Smoothing of functional data

Smoothing is a standard technique used in FDA. If we assume that our functional data $Y(t)$ is observed through the model: $Y(t_j) = X(t_j) + \varepsilon(t_j)$ for $j = 1, \dots, m$, where the residuals $\varepsilon(t_j)$ are independent from $X(t_j)$, we can get back the original signal $X(t_j)$ using a linear smoother,

$$\hat{X}(t_j) = \sum_{i=1}^{m} s_j(t_i) Y(t_i)$$

where $s_j(t_i)$ weights the observation $Y(t_i)$ for fitting the value of X at the time t_j. If $\mathbf{S}$ is the $m \times m$ matrix that $(\mathbf{S})_{i,j} = s_i(t_j)$, then the linear smoothing can be rewritten in matrix terms,

$$\hat{\mathbf{X}} = \mathbf{SY}$$

Two procedures have been considered. The first one is the representation in a $\mathcal{L}_2$ basis (or penalized basis) and the second one is based on smoothing kernel methods.

- Finite representation in a basis. A curve can be represented by a basis when the data is assumed to belong to $\mathcal{L}_2$ space. It can be done using equation (1.1) and the projection (or smoothing) matrix is given by: $\mathbf{S} = \Psi(\Psi^\top \Psi)^{-1}\Psi^\top$, with degrees of freedom of the fit $k_n = trace(\mathbf{S})$. If smoothing penalization is required, a parameter λ will also be provided and in this case, the projection matrix $\mathbf{S}$ is: $\mathbf{S} = \Psi(\Psi^\top \Psi + \lambda R)^{-1}\Psi^\top$, where R is the roughness penalty matrix. The most common roughness penalty is the integral of the square of the second derivative.

- Kernel smoothing. In its simplest form the smoothing of matrix $\mathbf{S}$ is given by Nadaraya–Watson estimator: $s_i(t_j) = K\left(\frac{t_i - t_j}{h}\right) / \sum_{k=1}^{m} K\left(\frac{t_i - t_k}{h}\right)$ where $K(\cdot)$ is the kernel function. Other possibilities for $\mathbf{S}$ are the k-nearest neighbors estimator or the local linear regression estimator, see Ferraty and Vieu (2006) for further details. Many different types of kernels can be considered.

1.3.3 Distance correlation

Distance correlation $\mathcal{R}$ is a measure of dependence between random vectors introduced by Székely et al. (2007). The distance correlation satisfies $0 \leq \mathcal{R}(X,Y) \leq 1$ and its interpretation is similar to the squared Pearson's correlation. However, the advantages of distance correlation over the Pearson correlation is that it defines $\mathcal{R}(X,Y)$ for arbitrary finite dimensions of X and Y and $\mathcal{R}$ characterizes independence, i.e. $\mathcal{R}(X,Y) = 0 \Leftrightarrow X,Y$ are independent. Recently, Lyons (2013) provided conditions for the application of the distance correlation to Hilbert spaces.

Distance correlation is a tool applied in this document to:

- Avoid collinearity. In Chapter 3 distance correlation $\mathcal{R}$ is useful for designing a functional linear model (avoiding variates with high collinearity).

- Avoid concurvity. In functional context, concurvity occurs when certain relationship among the functions holds. The distance correlation could identify information in one distance that is similar to another (for instance, in functional additive models), see Section A.3.

- Depth variable selection. In Chapter 4, distance correlation is used to select the useful depths whilst attempting to maintain a low dimension of the classification problem (covariates used in the model). Thus, the distance correlation between the multivariate vector of depths and the indicator of the classes is computed and the depth that maximizes the distance correlation is selected.

- Impact points selection. In Chapter 5, we study the utility of distance correlation as a method for impact point selection restricted to a single functional covariate.

- Estimate non-linear relations. $\mathcal{R}$ characterizes independence and so, can also detect relationships, other than linear, among variables.

- Variable selection for functional regression. Chapter 6 applies our proposal as a method for variable selection (from different nature, categorical, multivariate and functional) in which the distance correlation is computed among the residuals of the current model with each candidate covariate.

1.3.4 Data depths for functional data

Exploratory techniques for the FDA can help the user to analyse datasets by identifying features, detecting possible errors or unexpected variability and summarizing data. A systematic review of the literature on the notions of depth allows us to know how deep a data point is in the sample. An useful tool for exploratory techniques in a multivariate framework is the notion of depth that measures how deep (in some sense representative) is a data point respect to a sample. Several definitions of depth can be provided for FDA and Chapter 2 includes those contained in the works of Cuevas et al. (2007) and Febrero et al. (2007) along with other proposals.

- *Fraiman and Muniz depth (FM).* FM depth was the first depth proposed in a functional context, see Fraiman and Muniz (2001).
  - *FM^p depth*, is an adaptation of *FM* depth for multivariate functional data. A p-summarized version of FM–depth is proposed assuming all the data have the same support (this happens, for instance, when using the curves and their derivatives).
- *h–Mode depth (hM).* Cuevas et al. (2007) proposed the hM depth (also called mode depth) as a functional generalization of the likelihood depth to measure how surrounded one curve is with respect to the others.
  - *hM^p*, is an adaptation of hM for multivariate functional data. The method constructs a metric that combines the p-dimensional metric like, for example, the Euclidean metric. It is recommended that the different metrics have similar scales to prevent one single component from dominating the overall distance.
- *Random Projection Methods.* Several depths based on random projections basically use the same scheme, see Cuesta-Albertos et al. (2007) with the main difference in the way it summarizes the information:
  - *Random Projection (RP).* Proposed in Cuevas et al. (2007), it uses univariate half space (HS) depth and summarizes the depths of the projections through the mean.
  - *Random Half Space (RT).* Proposed in Cuesta-Albertos and Nieto-Reyes (2008), also uses the HS depth in both steps: as the univariate depth to compute the depth for every projection and to summarize the projections, now using the minimum value.
  - *Random Projection with derivatives (RPD).* Also proposed in Cuevas et al. (2007), it is an example on the extension to multivariate functional data combining the original curves and its derivatives. This is done using a p-variate depth with the projections.
- *FM^w, RP^w and hM^w depth.* For p-dimension multivariate functional data, a weighted version is defined by combining the functional univariate depth computed for every p component.
- *Spatial depth (SD).* In the two proposals in Sguera et al. (2014), the *Functional Spatial Depth (FSD)* provides results that are very similar to *FM* or *RP* and the *Kernelized Functional Spatial Depth (KFSD)* behaves like the *hM* depth.

Some authors have proposed other functional depth measures over the last years but they are all closely related with the three above-mentioned depths. For instance, the *Modified Band Depth (MBD)* proposed in López-Pintado and Romo (2009) can be regarded as a particular case of the *FM* depth using simplicial depth as a univariate depth. The works by Ieva and Paganoni (2013) and Claeskens et al. (2014) are in line with the extension of the *FM* depth to multivariate functional data with common support.

1.4 Functional depth classification models

The models of supervised classification have been based on two main lines: as an adaptation of regression models and as a result of the distributional characteristics of the population. Depth methods are substitutes for the latter.

The proposed classifier (DD^G–classifier) provides some extensions to cover multivariate functional data, combining the information of all the components and reducing their dimension without risking the loss of relevant information for the classification problem. In particular, Section 4.2.1 considers several proposed depths used in the DD^G–classifier procedure to select a suitable classification rule. We now have a purely multivariate classification problem in dimension G, which many procedures successfully handle based on either discriminant or regression ideas (see, for example, Ripley (1996)). We have chosen to use the following (see Section 4.2.2 for more details): based on discriminant analysis like the linear and quadratic discriminant analysis, based on logistic regression models like the generalized linear models (GLM) and the generalized additive models (GAM) and based on non-parametric estimates of the group densities like the k–nearest neighbour (kNN) and the Nadaraya–Watson estimator.

1.5 Functional regression models (FRM)

A regression model is said to be "functional" when at least one of the variables (either a predictor variable or response variable) is functional. One of the main contributions of this thesis is the implementation of several functional regression models (FRM) in which the response, or at least one of the covariates, has a functional nature. FRM have been widely studied, especially when the response is scalar and the covariates are functional. Linear models are treated in Ramsay and Silverman (2005), whereas Ferraty and Vieu (2006) is mainly devoted to nonlinear ones. Müller and Stadtmüller (2005) extended generalized linear models (GLM) to the situation where some of the covariates are functional. Müller and Yao (2008) did the same for the additive model (FAM), and Febrero-Bande and González-Manteiga (2013) proposed the generalized additive model (FGKAM).

First of all, we present a general approach from which these different FRMs will be deduced. Let $Y \in \mathbb{R}$ be the scalar response, $\mathbf{Z} \in \mathbb{R}^p$ the vector covariate and $\mathcal{X} = \{X^j(t)\}_{j=1}^q$ the q functional covariates. A general model is:

$$Y = m(\mathbf{Z}, \mathcal{X}) + \varepsilon, \tag{1.2}$$

where $m(\cdot)$ is an unknown function that combines the covariates and ε is an error process. Below, we relate the previous model with some important models introduced in literature over the last few years.

1.5.1 Functional linear regression (FLR)

The first proposal assumes linearity in the conditional expectation of the response and covariates. This is a classical functional linear model (FLR), that for the case of scalar response Y and a single functional covariate $\mathcal{X} = X(t)$ can be written as:

$$\mathbf{E}[Y|\mathcal{X}] = \alpha + \langle \beta, \mathcal{X} \rangle \tag{1.3}$$

where α is the intercept parameter and β is the functional parameter of interest to be estimated. The following is a usual representation of FLM when β and $\mathcal{X}$ belong to the $\mathcal{L}_2(T)$ Hilbert space in $T = [a,b]$ with $\langle \cdot, \cdot \rangle$ denoting the inner product:

$$Y = \alpha + \langle \beta, \mathcal{X} \rangle + \varepsilon = \alpha + \int_a^b \beta(t) X(t) dt + \varepsilon \tag{1.4}$$

where $\beta = \beta(t)$ is the unknown functional parameter and ε is the error of mean 0.

The main idea is the projection of each pair $X(t)$ and $\beta(t)$ onto a finite number of elements of a functional basis. They can either be chosen fixed in advance such as the B-spline, Fourier or Wavelet basis (see for example; Cardot et al. (2003) and Ramsay and Silverman (2005)) or PC or PLS basis (see for example; Cardot et al. (1999) and Aguilera et al. (2010)).

The FLR implemented with q functional covariates $\left\{\mathcal{X}^j\right\}_{j=1}^q$ is,

$$Y = \alpha + \sum_{j=1}^{q} \left\langle \beta_j, \mathcal{X}^j \right\rangle + \varepsilon = \alpha + \sum_{j=1}^{q} \int_{a_j}^{b_j} X^j(t) \beta_j(t) dt + \varepsilon, \tag{1.5}$$

where the functional covariates can be measured along different intervals.

The FLR models in Equations (1.4) and (1.5) have been widely studied in the literature and the functional coefficients may be estimated in a different way, as can be seen in Chapter 2. The notation $\langle \cdot, \cdot \rangle$ suggests that the extension to incorporate scalar covariates $\left\{Z^j\right\}_{j=1}^p$, in this model and the classical multivariate case can be done the same way.

1.5.2 Functional additive regression (FAR)

The model given by Equation (1.5) involves some functional predictors, but, in an inflexible way because it assumes linear relationship between the predictors and the response. Just as in the context of the standard regression with scalar covariates, more accurate fits may be obtained by modelling a non–linear relationship. In this case, a more flexible regression model is given by:

$$Y = \sum_{j=1}^{q} m_j(\mathcal{X}^j) + \varepsilon, \tag{1.6}$$

When $q = 1$ we have the non–parametric regression model with functional covariate widely considered in the literature, as can be seen in the book of Ferraty and Vieu (2006) and related papers. The estimation for $m(\cdot)$ is given using different procedures based on smoothing techniques such as splines and kernel smoothing (Nadaraya–Watson, Local Linear and kNN estimators), see Section 2.3.4.

1.5.3 Semi-functional linear models (SFLM)

Exogenous scalar variables with the additional information given by the functional covariates are often present. The semi–functional linear regression introduced and studied in Aneiros-Pérez and Vieu (2006) generalizes the non–parametric model with a single functional covariate $\mathcal{X}$ incorporating a linear component with p exogenous scalar variables $\mathbf{Z} = \left\{Z^1, \ldots, Z^p\right\}$ in the regression function:

$$Y = \alpha + \mathbf{Z}^\top \gamma + m(\mathcal{X}) + \varepsilon, \tag{1.7}$$

where γ are the p unknown parameters, see Section 2.3.5 for more details on model implementation.

1.6 Extensions of functional regression models

This section introduces models widely treated in the literature. The Appendix A collects further details on the models presented below.

1.6.1 Functional generalized linear regression models (FGLM)

In several applications the functional linear model (FLM) may be too restrictive, for instance when the response is binary or a count. One natural extension of FLMs is the functional generalized linear regression model (FGLM) proposed by Müller and Stadtmüller (2005), which allows various types of response. Its expected value is related to this linear predictor via a link function. The GLM framework generally assumes that scalar response can be chosen within the set of distributions belonging to the exponential family. This general approach includes the case of the FLM, functional Poisson and binomial regression. The latter leads to procedures for classification and discrimination of stochastic processes and functional data.

Let $Y \in \mathbb{R}$ be the scalar response, $\mathbf{Z} \in \mathbb{R}^p$ the vector covariate, and $\{\mathcal{X}^j\}_{j=1}^q$ the functional covariate with values in the product of q infinite-dimensional Hilbert spaces, the FGLM has the following expression:

$$\mathbf{E}[Y|\mathbf{Z},\mathcal{X}] = g^{-1}\left(\alpha + \langle \mathbf{Z},\gamma\rangle + \sum_{j=1}^{q}\left\langle \mathcal{X}^j,\beta_j\right\rangle\right) = g^{-1}\left(\alpha + \mathbf{Z}^{\top}\gamma + \sum_{j=1}^{q}\int_{a_j}^{b_j} X^j(t)\beta_j(t)dt\right) \tag{1.8}$$

where $g(\cdot)$ is the link function, α is the intercept, γ contains the p regression coefficients using the inner product in Euclidean vector space and $\beta(t)$ contains the functional regression coefficients using the inner product in Hilbert space.

Appendix A.2 illustrates with an R example the use of the GLM model in the functional case.

1.6.2 Functional generalized spectral additive regression models (FGSAM)

We implement the functional generalized additive model (FGAM). The model can be regarded as the natural extension of GAM models with functional predictors or as the additive version of the FGLM seen in Equation 1.8. Regarding the latter, the FGAM model estimates non-linear relations between the response and the predictors using the smooth functions $r_k(\cdot)$ and $m_j(\cdot)$ for scalar and functional covariates, respectively.

$$\mathbf{E}[Y|\mathbf{Z},\mathcal{X}] = g^{-1}\left(\sum_{k=1}^{p} r_k(Z^k) + \sum_{j=1}^{q} m_j(\mathcal{X}^j(t))\right) \tag{1.9}$$

The estimation of the functions $m_j(\cdot)$ can be done in two ways: following Müller and Yao (2008) using the spectral decomposition of the $X(t)$ (in the so-called functional generalized spectral additive model (FGSAM)) or following Febrero-Bande and González-Manteiga (2013) using functional kernel for the estimation of $m_j(\cdot)$ functions (in the so-called functional generalized kernel additive model (FGKAM)). We develop the FGSAM and FGKAM procedures in `fda.usc` package, and the Appendix A.3 has an example of their usage.

1.6.3 Functional regression models with dependent errors

All previous models assume independent errors. However, the observations commonly present a structure of dependence that should be taken into account. The GLS model allows the joint estimation of the parameters of the model and the error term for the case of scalar response and multivariate covariates (Kariya and Kurata, 2004). For multivariate and functional covariates, the model (called FGLS) states that,

$$Y = \alpha + \langle \mathbf{Z}, \gamma \rangle + \langle \mathcal{X}, \beta \rangle + \varepsilon = \alpha + \mathbf{Z}^{\top}\gamma + \int_a^b X(t)\beta(t)dt + \varepsilon \tag{1.10}$$

where ε is now a random vector with mean 0 and covariance matrix $\Omega = \mathbf{E}[\varepsilon\varepsilon']$. This model includes many other models as its special cases, all of which are based on $\Omega = \Omega(\phi) = \sigma^2\Sigma(\phi)$, where ϕ is the parameter associated with the dependence structure of Ω.

Our proposal (see Chapter 3) extends the classical theory of Kariya and Kurata (2004) to the functional case by adapting the GLS criterion. The extension of the GCV criterion proposed by Carmack et al. (2012) is also considered. In practice, our implementation calls the `gls` function of *nlme* package. Therefore, we observe that the equations for prediction, programmed by the original authors (Pinheiro *et al.* (2014)) of the package *nlme*, do not take the Σ parameter estimation into account (see Section 3.2).

To alleviate the computational burden, we develop a new proposal that consists in separating the estimation of the dependence structure from the parameters associated to the regression in an iterative way (called iGLS). The iGLS, see Section 3.2, is proven to be equivalent to classical GLS (see, for instance, Goldstein (1986)).

We implement functions that estimate and predict the functional regression model with correlated errors, and also using the iterative scheme (iGLS). For the latter, we have developed the following two simple structures to Σ for fit serial dependence structure:

- In iGLS-AR(p) scheme, the procedure automatically fits the autoregressive order p of the errors in each iteration.

- In iGLS-ARMA(p,q) scheme, the user must specify the parameters p and q of the autoregressive–moving–average (ARMA(p,q)) model, which fits the serial error dependence.

1.6.4 Functional regression models with functional response (FRM.FR)

This section is devoted to functional regression models with a functional response variable (FRM.FR) and at least one functional covariate. Several authors have analyzed the case of functional response models e.g., Faraway (1997), Chiou et al. (2004), Ferraty et al. (2012) and Ramsay and Silverman (2005).

We follow the proposal made by Ramsay and Silverman (2005) that models the relationship between the functional response, now $\mathcal{Y} = Y(s) \subset \mathcal{L}_2(S)$ and the functional covariate $\mathcal{X} = X(t) \subset \mathcal{L}_2(T)$ in the following way:

$$Y(s) = \alpha(s) + \langle \beta, \mathcal{X} \rangle + \varepsilon(s) = \alpha(s) + \int_a^b \beta(t,s)X(t)dt + \varepsilon(s) \tag{1.11}$$

Now, $\alpha(s)$ and $\varepsilon(s)$ are functions (intercept and error term respectively), and the regression parameter $\beta(t,s) \in \mathcal{L}_2(T \times S)$ is a bivariate function. A function in `fda.usc` package carries out a FRM.FR, where both, dependent and independent variable, are functional. When time $s = t + 1$, it is the case of functional autoregressive model of order one. The estimation theory for this and other Hilbertian processes in the framework of functional time series is developed in the works of Bosq (2000) and Horvath and Kokoszka (2012).

1.7 Feature selection for functional data

Techniques for variable selection have been widely studied in statistics. There are only a few studies that are framed within the functional framework; Ferraty and Vieu (2009) proposed a version of boosting applied to the functional context, Ferraty *et al.* (2010) proposed treating the selection of the optimal points of a functional data in a regression environment, and Berrendero *et al.* (2015) studied the selection of optimal points of a functional data for classification.

1.7.1 Impact points in functional regression models

Our aim is to select significant impact points (also called most predictive design points) for a regression model with a functional covariate and scalar response. In Chapter 5 we combine several useful statistical techniques and ideas properly done in our final algorithm. Section 5.2.1 provides a brief summary of the functional parametric model (see Equation 1.4) and non-parametric regression models (see Equation 1.6). Both are used to estimate leaf water content from reflectance. However, one drawback of these kinds of methods is that the output is not directly interpretable in terms of original variables. Hence the great interest in variable selection methods, especially in those where the output only depends on the data, rather than any underlying modelling. Following this line, Berrendero *et al.* (2015) extends the minimum Redundancy Maximum Relevance (mRMR) procedure proposed by Peng et al. (2005) to select the relevant points of $X(t)$. A different solution was proposed in Zhao et al. (2012), using a wavelet-based LASSO procedure. More recently, this approach was improved by means of screening and penalty factor weighting schemes (Zhao et al., 2015).

Let m be potential predictors, which compose a functional covariate $\{X(t_j)\}_{j=1}^{m}$. The goal is to select a small subset S of impact points $\{X(t_l)\}_{l=1}^{k}$, with $k \leq m$ which are the most informative variables for predicting the response. For this purpose, we use the distance correlation proposed by Székely et al. (2007) for computing the local maxima of distance correlation (LMDC) following the same ideas proposed in (Berrendero *et al.*, 2016) for impact points selection in functional classification framework. The main advantage of our approach is that it is an incremental rule, and the list of candidate covariate is ordered by the LMDC values. This requires neither projection nor transformation of the impact points.

1.7.2 Variable selection in functional regression

The variable selection problem, in a general regression model, tries to find the subset of covariates that best predicts or explains a response. Our purpose is to provide an automatic procedure for selecting regression models with a subset of the available covariates of different nature (mainly functional, scalar and categorical).

The stepwise regression is the most widely-used model selection technique throughout the classical papers by Akaike (1973), Mallows (1973), Schwarz and other (1978) and Stone (1979). The work by Tibshirani (1996) proposes the LASSO estimator that includes a $\mathcal{L}_1$-type constraint to obtain the optimal subset of covariates. All the previous solutions are not completely satisfactory in a functional data framework, especially when the number of possible covariates can be arbitrarily large.

Our aim is to select significant covariates for a functional additive regression (FAR) model (see generalized version in Equation 1.9) with scalar response. The number of variates can be extraordinarily large, so we construct the regression model sequentially, i.e. from the trivial model up to the one that includes all the useful information provided by the covariates in the subset of covariates S. To do this, we use the distance correlation between the residual of the model and the potential covariate.

References

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems,*, 104(2), 289-305.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

Aneiros-Pérez, G. and Vieu, P. (2006). Semi-functional partial linear regression. *Statistics & Probability Letters*, 76(11):1102–1110.

Antoniadis, A. and Sapatinas, T. (2003). Wavelet methods for continuous-time prediction using hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, 87(1):133–158.

Berrendero, J. R., Cuevas, A., and Torrecilla, J. L. (2015). Variable selection in functional data classification: A maxima-hunting proposal. *Statistica Sinica*, 26.

Berrendero, J. R., Cuevas, A., and Torrecilla, J. L. (2016). The mrmr variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation*, 86(5):891–907.

Bosq, D. (2000). Linear processes in function spaces: Theory and applications, volume 149 of lecture notes in statistics.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–592.

Carmack, P. S., Spence, J. S., and Schucany, W. R. (2012). Generalised correlated cross-validation. *Journal of Nonparametric Statistics*, 24(2):269–282.

Chiou, J.-M., Müller, H.-G., and Wang, J.-L. (2004). Functional response models. *Statistica Sinica*, pages 675–693.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505):411–423.

Cuesta-Albertos, J. and Nieto-Reyes, A. (2008). The random tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988.

Cuesta-Albertos, J. A., Febrero-Bande, M., and Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *Test*, 26(1):119–142

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2007). A sharp form of the cramer–wold theorem. *Journal of Theoretical Probability*, 20(2):201–209.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Cuevas, A., Febrero-Bande, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

Damon, J. and Guillas, S. (2005). Estimation and simulation of autoregressive hilbertian processes with exogenous variables. *Statistical inference for stochastic processes*, 8(2):185–204.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239.

Escabias, M., Aguilera, A., and Valderrama, M. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3):254–261.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2007). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, 19(4):331–345.

Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *Test*, 22(2):278–292.

Febrero–Bande, M., González–Manteiga, W. and Oviedo de la Fuente M. (2017). Variable selection in Functional Additive Regression Models. *In Functional Statistics and Related Fields*, 113-122. Springer, Cham.

Febrero-Bande, M., González-Manteiga, W., and Oviedo de la Fuente M. (2018). Variable selection in Functional Additive Regression Models. Published online in *Computational Statistics*, DOI: `https://doi.org/10.1007/s00180-018-0844-5`.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda. usc. *Journal of Statistical Software*, 51(4):1–28.

Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2011). Recent advances on functional additive regression. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 97–102. Springer.

Ferraty, F., Hall, P., and Vieu, P. (2010). Most predictive design points for functional data predictors. *Biometrika*, 94(4):807–824.

Ferraty, F. and Romain, Y. (2010). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press Oxford.

Ferraty, F., Van Keilegom, I., and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109:10–28.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer.

Ferraty, F. and Vieu, P. (2009). Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis*, 53(4):1400–1413.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1):43–56.

Horvath, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. Springer Science & Business Media

Ieva, F. and Paganoni, A. M. (2013). Depth measures for multivariate functional data. *Comm. Statist.-Theory Methods*, 42(7):1265–1276.

Kariya, T. and Kurata, H. (2004). *Generalized least squares*. Wiley.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). *dd*–classifier: Nonparametric classification procedure based on *dd*–plot. *Journal of the American Statistical Association*, 107(498):737–753.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805.

Müller, H. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Ordóñez, C., Oviedo de la Fuente, M., Roca-Pardiñas, J., and Rodríguez-Pérez, J. R. (2018). Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. *Chemometrics and Intelligent Laboratory Systems*, 173:41–50.

Oviedo de la Fuente, M., Febrero-Bande, M., Muñoz, M. P., and Domínguez, À. (2018). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PloS one*, 13(4):e0194250.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Pinheiro J, Bates D, DebRoy S, Sarkar D. R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. Available at https://cran.r-project.org/web/packages/nlme. 2014.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge Uni. Press, Cambridge.

Schwarz, G. and other (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *Test*, 23(4):725–750.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):276–278.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Yenigün, C. D. and Rizzo, M. L. (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, 85(8):1692–1705.

Zhao, Y., Chen, H., and Ogden, R. T. (2015). Wavelet-based weighted lasso and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics*, 24(3):655–675.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

Chapter 2

Statistical computing in functional data analysis: The R package `fda.usc`

Reference paper

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4), 1–28. DOI: `http://dx.doi.org/10.18637/jss.v051.i04`.

Chapter 3

Predicting seasonal influenza transmission using functional regression models with temporal dependence

Reference paper

Oviedo de la Fuente, M., Febrero-Bande, M., Muñoz, M. P., and Domínguez, A. (2016). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PloS one* 13(4), e0194250, DOI: `https://doi.org/10.1371/journal.pone.0194250`

Chapter 4

The DD^G-classifier in the functional setting

Reference paper

Cuesta–Albertos, J. A., Febrero–Bande, M., and Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *TEST*, 26(1), 119-142.
DOI: `https://doi.org/10.1007/s11749-016-0502-6`.

Chapter 5

Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach

Reference paper

Ordóñez, C., Oviedo de la Fuente, M., Roca-Pardiñas, J., & Rodríguez-Pérez, J. R. (2017). Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. *Chemometrics and Intelligent Laboratory Systems.* DOI: https://doi.org/10.1016/j.chemolab.2017.12.001.

Chapter 6

Variable selection in functional additive regression models

Reference paper

Febrero-Bande, M., González-Manteiga, W., and Oviedo de la Fuente M. (2018). Variable selection in Functional Additive Regression Models. https://doi.org/10.1007/s00180-018-0844-5.

Chapter 7

Conclusions and future research

Contents

7.1 Conclusions

The main achievements of the thesis are the following:

i. Implement in R an integrated framework (package *fda.usc)* where many of the techniques of FDA are available. Below are some of the most prominent techniques for exploratory FDA along the different chapters of the thesis:

   (a) FD representation (fixed and data-driven basis).

   (b) Smoothing for FD.

   (c) Computation of most relevant metrics and semimetrics for FD.

   (d) Depth measures for (multivariate) functional data.

   A general objective has been to develop, maintain, update and improve the *fda.usc* library that is included in the CRAN task view:
   `https://cran.r-project.org/web/views/FunctionalData.html`.

ii. Develop a general framework to extend the Functional Regression Models to the case of a response from the exponential family. Appendix of Chapter 2 extends the GLM and GAM models to the functional case.

iii. Extend that functional linear model (FLM) model to analyse functional data that have a temporal or spatial dependence.

iv. Chapter 3 extends the ideas of generalized least squares (GLS) methods to functional regression models (FRM) with scalar response.

v. Classifiers based on functional regression models have been developed by different classification schemes (like majority voting) and for the case of multi-class response.

vi. A new classification proposal based on depths was done in Chapter 4 and compared with classical classification techniques for FD.

vii. An algorithm that selects the most relevant points of impact in a functional regression problem is proposed in Chapter 5. This proposal uses distance correlation as an alternative to the work of Ferraty *et al.* (2010).

viii. Chapter 6 proposes a general framework for the case of variable selection in regression with functional, scalar and categorical covariates. To the best of our knowledge, no previous works provide such a general approach like the one proposed here. The proposal is restricted to addtive models by their balanced compromise among predictive ability and simplicity even for a large number of covariates.

7.2 Future research

This chapter presents some ideas for future contributions in different areas related with the scope of the thesis. Some of them were already pointed out in the conclusions for each chapter, but the most remarkable ones are summarized in the following:

i. **Complex data analysis**. Propose statistical techniques for functional data in two dimensions (2D) and for multivariate functional data.

   In Chapter 2, the following research lines are still open.

   (a) **Functional data in 2D**. This thesis considers functional data analysis on 1D dimensional domain. The extension to 2D can present similar problems such as: representation, regularization, visualization and smoothing.

   (b) **Multivariate Functional Data**. The multivariate functional data can be treated as realizations of multivariate random processes, see Hubert et al. (2015). Outlier detection methods developed in Chapter 2 can be extended, for instance, to the use of depths for multivariate functional outliers. Multivariate Functional Data is treated in Chapter 5.

   (c) **Multivariate Functional k-means**. Similarly to the previous procedure, a depth-based method can be provided for multivariate functional data.

   (d) **Functional Response Models (FRM)**. Mostly all functional regression models contained in this work are devoted to the case of scalar response and it could be interesting to extend them to functional response case. This could be done by following the works of Faraway (1997), Ramsay and Silverman (2005) and Chiou et al. (2004) and Ferraty *et al.* (2012).

   (e) **Functional Time Series**. The treatment of functional time series can partially follow the developments of functional regression models with functional response.

ii. **FGLS models**. Our method, presented in Chapter 3, can additionally be used to explore more complex dependence structures like heterogeneous covariances by groups or even spatio–temporal modelling. An open problem is how to relax the linearity assumption of the FGLS to non-linear assumption. Another extension intends to establish the functional mixed model proposed by Scheipl et al. (2016) as a general framework allowing, for example:

   (a) **Random effects**. Functional random effects with flexible correlation structures for, e.g., spatial, temporal. Include functional spatial procedures implemented by Delicado et al. (2010).

   (b) **Functional Quantile Regression**. Extend the works of Koenker and Hallock (2001), to functional case extending the work of Kato et al. (2012).

(c) **GAMLSS Models**. Extend the generalized additive models for location, shape and scale (GAMLSS) for functional data, see Stasinopoulos et al. (2007).

iii. **Functional Supervised Classification**. Adapt the classification proposal to the multiclass problem, the problem of unbalanced groups, multivariate functional data and complex data (like the case of hyperspectral image).

iv. **Efficient R programming.** Improve the computational efficiency of the routines implemented in *fda.usc* both in time and in consumed memory (using parallelization for example or adapting the code so that it can handle large volumes of information).

References

Chiou, J.-M., Müller, H.-G., and Wang, J.-L. (2004). Functional response models. *Statistica Sinica*, pages 675–693.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3):254–261.

Ferraty, F., Vieu, P., and Van Keilegom, I. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109: 10 – 28.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202.

Kato, K. et al. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Scheipl, F., Gertheiss, J., Greven, S., et al. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, 10(1):1455–1492.

Stasinopoulos, D. M., Rigby, R. A., et al. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

Appendices

Appendix A

Supplement to Chapter 2

This supplement is organized as follows. Appendix A.1 shows the main functions implemented corresponding to the methods presented throughout Chapter 2 and in this appendix. The `fda.usc` package documentation contains detailed information about the use of these functions. Appendix A.2 contains some of the main contributions developed from the publication of paper by Febrero–Bande and Oviedo de la Fuente (2012) up to the present. We focus on functional regression models, in particular those in which response belongs to exponential family, the response is functional or the relationship between response and predictors is non-linear.

A.1 Functions added to *fda.usc* package

This section shows the R functions used in Chapter 2 included in `fda.usc` package. Table A.1 contains functions for fdata class definition, Table A.2 contains the principal functions for functional exploratory analysis, and Table A.3 shows those functions for functional regression.

R Function	**fdata class objects**
fdata()	converts object `fd`, `fds`, `sfts`, `vector`, `matrix` and `data.frame` to `fdata`
fdata2fd()	converts a `fdata` object to a `fd` object (fda).
fdata.deriv()	computes the derivative of a trajectory
plot.fdata()	plots `fdata` class object
	Basic operations on `fdata` objects
abs, **sqrt**, **floor**,	Group Math I
trunc, **round**, **exp**,	Group Math II
log, **cos**, **sin**	Group Math III
all, **any**, **sum**, **prod**,	Group Summary I
min, **max**, **range**	Group Summary II
$==$, $+$, $-$, $*$, $/$	Aritmetric
[], , **is.fdata**, **anyNA**,	S3 methods I
c, **dim**, **ncol**, **nrow**	S3 methods II
count.na, **order.fdata**	S3 methods III
	Functional data Manipulation
fdata2pc()	Penalized PC for functional data
fdata2pls()	Penalized PLS for functional data

Table A.1: R functions for `fdata` class.

R Function	**Functional smoothing**
S.basis()	Smoothing matrix by fixed basis
S.LLR()	Smoothing matrix by Local Linear Regression
S.NW()	Smoothing matrix by Nadaraya-Watson kernel estimator
S.KNN()	Smoothing matrix by K nearest neighbors estimator
min.basis()	Smooth functional data by basis representation
min.np()	Smooth functional data with a kernel bandwidth selection
CV.S()	Leave-one-out cross-validation (CV) score
GCV.S()	Generalized correlated cross-validation (GCV) score
	Distance between functional elements and related functions
inprod.fdata()	Inner product of `fdata`
norm.fdata()	$\mathcal{L}_p$-norm for `fdata`
metric.lp()	$\mathcal{L}_p$-metric for functional data
metric.hausdorff()	Hausdorff distances between two sets of curves
metric.kl()	Kullback–Leibler distance between two groups of densities
metric.dist()	Wrapper function of stats:::dist() function
semimetric.basis()	Semi-metric distances based on basis
semimetric.deriv()	$\mathcal{L}_2$ metric between derivatives of the curves based on B-spline
semimetric.fourier()	$\mathcal{L}_2$ metric between the curves based on Fourier
semimetric.mplsr()	Distance between curves based on the PLS
semimetric.pca()	Distance between curves based on PC
	Functional univariate data depth
depth.FM()	Computes Fraiman and Muniz (FM) depth
depth.mode()	Computes modal depth
depth.RT()	Computes random tukey (RT) depth
depth.RP()	Computes random project (RP) depth
depth.RPD()	Double random project depth (RPD) depth
	Functional outlier detection
outliers.thres.lrt()	Functional outliers by LRT
outliers.depth.trim()	Functional outliers using trimmed data
outliers.depth.pond()	Functional outliers using weighted data

Table A.2: R functions for functional exploratory analysis.

R Function	**Functional linear models (FLR)**
fregre.lm()	FLR for functional (and non functional) covariates using basis
fregre.basis()	FLR for functional predictor using basis
fregre.basis.cv()	FLR using selection of size of a basis
fregre.pc()	FLR (ridge or penalized) using PCA
fregre.pc.cv()	Optimal selection of PC's for FLR (ridge or penalized)
fregre.pls()	FLR using penalized PLS
fregre.pls.cv()	Optimal selection of PLS's for FLR (ridge or penalized)
	Model fitting
summary()	Summarizes information from fitted models
predict()	Predictions from object fitted
influence.fdata()	Computes influence measures from FLM
fregre.bootstrap()	Gives diagnostic for parameters derived from a FLM
	Functional non-linear models FNLR
fregre.np()	FNLR using kernel estimation
fregre.np.cv()	Optimal choice of bandwidth for a FNLR
fregre.plm()	Functional semi-linear model
	Generalized regression models
fregre.glm()	Functional GLM model
fregre.gsam()	Functional Spectral GAM model
fregre.gkam()	Functional Kernel GAM model
	Functional response model (FRM)
fregre.basis.fr()	Fits FRM using basis representation

Table A.3: R functions for functional regression.

See Appendix of Chapter 4 for classification methods based on regression. The fda.usc package contains other relevant functions and shortcuts not presented in this thesis.

A.2 Functional generalized regression models (FGLM): R example

We develop the `fregre.glm()` function to fit the FGLM models (defined in equation 1.8) in R. The function extends the GLM model to the functional case and generalizes the FLM model to several types of response. This is the reason why the main arguments of the function combine those of the functions `glm()` and `fregre.lm()`. The function has been implemented following the structure of the "`glm`" class, so it is possible to use standard R tools such as the `summary()` and the `predict()` functions. In addition, specific features, such as the display of the estimated beta parameter, have been programmed.

Below, we show how to apply the FGLM model with binary response (dichotomized fat content, 1 for `fat>15`, 0 otherwise) in the Tecator dataset. In the following, we use a training sample (first 165 curves) of the second derivative of absorbance curves `X.d2` to estimate the response. As FLM, the main idea is to reduce the dimension of functional covariates to a few basis functions. The functional logistic regression is estimated by `fregre.glm()` that has the same arguments of the `fregre.lm()` function plus the family parameter (`family=binomial()`).

```r
data(tecator)
ind <- 1:165
Fat.bin <- ifelse(tecator[["y"]][,"Fat"] > 15, 1, 0)
```

```
dataf <- data.frame(tecator[["y"]][ind,], Fat.bin[ind])
names(dataf)[4] <- "Fat.bin"
X.d2 <- fdata.deriv(absorp, nderiv = 2)
ldata <- list("df" = dataf, "X.d2" = X.d2[ind])
basis.pc2 <- create.pc.basis(X.d2[ind], l = c(1))
basis.x <- list("X.d2"=basis.pc2)
res.pc <- fregre.glm(Fat.bin ~ X.d2, ldata, family = binomial,
basis.x = basis.x)
```

To illustrate this, the fitted object returned (`res.pc`) can be used in other functions of the (`glm`) class such as; `summary()`.

```
summary.glm(res.pc)
##
## Call:
## glm(formula = pf)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.49008  -0.07853  -0.01149   0.00303   2.01877
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.7509     0.7588   3.625 0.000289 ***
## X.d2.PC1    4072.5892   910.5123   4.473 7.72e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.713  on 164  degrees of freedom
## Residual deviance:  32.881  on 163  degrees of freedom
## AIC: 36.881
##
## Number of Fisher Scoring iterations: 9
```

If new data is observed, the response can be predicted by the `predict.fregre.glm()` function.

```
newx <- tecator[["absorp.fdata"]][-ind,]
newx.d2 <- fdata.deriv(newx, nderiv = 2, method = "fmm")
newdataf <- as.data.frame(newy)
newldata=list("df" = newdataf, "X.d2" = newx.d2)
pred.pc <- predict.fregre.glm(res.pc, newldata)
```

Logistic regression method can be used for a binary classification variable. The prediction of dichotomized fat content is correct by 98% in test sample (last 50 data). The confusion matrix between the theoretical response values (usually not observed) and the predicted values is,

```
y.pred1 <- ifelse(pred.pc > 0.5, 1, 0)
table(y.pred1, Fat.bin[-ind])
```

```
##
## y.pred1  0  1
##       0 23  1
##       1  0 26
```

Section D.2 provides some useful classification functions implemented in the `fda.usc` package.

A.3 Functional generalized additive regression models: R example

This section presents the functional generalized additive models (FGSAM), see Equation (1.9), that allows modelling non-linear relationships between a response variable (that belongs to the exponential family) and the predictor variables (functional or non-functional).

Müller and Yao (2008) proposed the generalized additive model (GAM) using a spectral decomposition of the $X(t)$, through the principal component (FPC) scores of $X(t)$. We develop FGSAM in `fregre.gsam()` function based on the `gam()` function of the *mgcv* package (Wood, 2006). The implemented function is not limited to data-driven basis as FPC or FPLS, but also allows basis representation such as: B-spline, Fourier or an ad-hoc basis chosen by the user. The choice of the type of basis and its number of components, as well as the type and degree of smoothing are crucial parameters to consider. Another proposal (Febrero–Bande and González–Manteiga, 2013) is estimate the smooth functions in a non-parametric way in the generalized kenel additive model (GKAM). For this implementation, a kernel is used for the estimation of $m()$ function. We develop FGKAM procedure in `fregre.gkam()` function. The procedure allows the metric and semi-metrics procedures used in the traditional functional non-parametric regression proposed in Ferraty and Vieu (2006): the function `fregre.np()` in *fda.usc* package.

Below is an example of logistic regression using the FGSAM model:

```
res.gsam <- fregre.gsam(Fat.bin ~ s(X.d2), family = binomial(),
data = ldata, basis.x = basis.x)
 summary(res.gsam)
##
## Family: binomial
## Link function: logit
##
## Formula:
## [1] "Fat.bin~+s(X.d2.PC1,k=-1)"
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.149      3.422   0.336    0.737
##
## Approximate significance of smooth terms:
##               edf Ref.df Chi.sq p-value
## s(X.d2.PC1) 1.393  1.691  12.42 0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.888   Deviance explained = 86.1%
## UBRE = -0.77875  Scale est. = 1         n = 165
```

Below is an example of logistic regression using the FGKAM model:

```
res.gkam <- fregre.gkam(Fat.bin ~ X.d2, family = binomial(),
data = ldata))
summary(res.gkam)
## ** Summary Functional Data Regression with backfiting algorithm **
##
## Family: binomial
## Link function: logit
##
## alpha= -1.02   n=  165
## Algorithm converged? Yes  Number  of iterations  7
##
## ****    ****    ****    ****    ****    ****
##                h cor(f(X),eta)  edf
## f(X.d2) 0.000412         0.975 66.1
## ****    ****    ****    ****    ****    ****
##
## edf: Equivalent degrees of freedom
## Residual deviance= 7.118  Null deviance= 227.713
## AIC=  141.399 Deviance explained= 96.9 %
## R-sq.= 0.98  R-sq.(adj)= 0.967
## Names of possible influence curves: No influence curves
```

Table A.2 includes the procedures described in Chapter 2, and Appendices A.2 and A.3 implemented in R.

A.4 Functional regression model with functional response

We follow the proposal made by Ramsay and Silverman (2005) which models the relationship between the functional response $Y(s)$ and the functional covariate $X(t)$ described in Equation (1.11) by basis representation of both. The `fregre.basis.fr()` carries out a FRM.FR, where both the predictor $X(t)$ (`x` argument) and the response $Y(s)$ (`y` argument) are functional. The implemented function allows objects of `fdata` class or directly covariates from `fd` class. The function also gives default values to `basis.t` and `basis.s` arguments to construct the `bifd` class object used in the estimation of $\beta(t,s)$. If `basis.s=NULL` or `basis.t=NULL` the function creates a b–spline basis by `create.bspline.basis()` function. The parameters of the function are:

```
args(fregre.basis.fr)
## function (x, y, basis.s = NULL, basis.t = NULL, lambda.s = 0,
##     lambda.t = 0, Lfdobj.s = vec2Lfd(c(0, 0), range.s),
##     Lfdobj.t = vec2Lfd(c(0,0), range.t), weights = NULL, ...)
```

Appendix B

Supplement to Chapter 3

This section shows R functions and a demo of its application (see Appendix B.1 and B.2 respectively). Appendix B.1 shows the main functions implemented corresponding to the methods presented throughout Oviedo de la Fuente et al. (2018). The *fda.usc* package contains detailed information about the use of functions of Table B.1.

B.1 Functions added to *fda.usc* package

R Functions	Models for Dependent Data
fregre.gls()	Fits functional generalized least squares (GLS) model
fregre.igls()	Fits functional generalized least squares (GLS) model iteratively
GCCV.S()	Generalized correlated cross-validation (GCCV) score
predict.fregre.gls()	Predictions from a functional `gls` object
predict.fregre.igls()	Predictions from a functional iterative `gls` object

Table B.1: R functions for FGLS regression

B.2 Functional GLS model: R example

Several data sets are used for illustration of methodological developments in FDA area, perhaps the Tecator data set is the one most used. However, it has gone unnoticed in previous works that the observations present a dependency on the registry. The following example adjusts a functional GLS model with a first-order autoregressive dependence AR(1) model on the errors.

To do this, we use the `fregre.gls()` function that has the same arguments as the `fregre.lm()` function to which the `correlation` argument have been incorporated. It has the same functionality as the argument with the same name of the function `gls()` where `corStruct()` object is expected describing the correlation structure.

```
data(tecator)
X.d2 <- fdata.deriv(tecator[["absorp.fdata"]], nderiv = 2)
ldata <- list("df"=tecator[["y"]], "X.d2" = X.d2)
res.gls <- fregre.gls(Fat ~ X.d2, data = ldata, correlation=corAR1())
coef(res.gls[["modelStruct"]], F)
## corStruct.Phi
##     0.4942661
```

The previous model is restricted to a structure determined by `gls()` function of *`nlme`* package. The function `fregre.igls()` is presented as an alternative because it allows any type of dependence structures designed by the user, see Section 3.2.

The code below shows a simple use of iterative scheme (iGLS). In particular, we use a iGLS-AR($p = 1$) scheme for error estimation.

```
res.igls=fregre.igls(Fat ~ X.d2, data = ldata,
correlation=list("cor.AR" = list()),control = list("order.max" = 1))
coef(res.igls[["corStruct"]][[1]])
##       ar1
## 0.4900419
```

Both examples estimate an AR(1) with $\phi \approx 0.49$. Thus, the estimation and the prediction made with these models will be more accurate than the classical functional models presented in Chapter 2 in which it is assumed that the errors are independent.

Appendix C

Supplement to Chapter 4

Appendix C.1 contains the list of the main functions developed or updated from the paper Cuesta-Albertos et al. (2017). Appendix C.2 shows the use in R of functional supervised classification methods.

C.1 Functions added to *fda.usc* package

R Function	**Functional Univariate Data Depth**
depth.FM()	Fraiman and Muniz (FM) depth
depth.mode()	Modal depth
depth.RT()	Random tukey (RT) depth
depth.RP()	Random project (RP) depth
depth.RPD()	Double random project depth (RPD) depth
	Functional Multivariate Data Depth
depth.FMp()	Fraiman-Muniz depth
depth.RPp()	Random Projections depth
depth.modep()	Modal depth
	Multivariate Data Depth
mdepth.MhD()	Mahalanobis depth
mdepth.HS()	Halfspace depth, also known as Tukey depth
mdepth.SD()	Simplicial depth
mdepth.LD()	Likelihood depth

Table C.1: Depth fda.usc functions for multivariate, functional and multivariate functional data

R Function	**Functional Classification based on Depths**
classif.depth()	Classification of functional data using maximum depth
classif.DD()	Classification procedure based on DD–plot
	Functional Classification based on regression models
classif.glm()	Classification using Functional Generalized Linear Models
classif.knn()	kNN Classifier
classif.kernel()	Nonparametric (Kernel) Classifier
classif.gsam()	Classification using Functional Generalized Spectral Additive Models
classif.gkam()	Classification using Functional Generalized Kernel Additive Models

Table C.2: fda.usc functions for functional classification

C.2 Functional supervised classification: R example

In supervised classification, we have a completely labelled sample by groups, i.e. $\{X_i(t), G_i\} \in \mathcal{F} \times \mathbb{G} = \{1, \ldots, G\}$ where G is a discrete variate that indicates the group that every observation belongs to. Our goal is to estimate the posterior probabilities for a new observation $X(t)$ to belong to each group (Bayes' rule), i.e.

$$p_g(X(t)) = \mathbf{P}(G = g | \mathcal{X} = X(t)) = \mathbf{E}\left[\mathbb{1}_{\{G=g\}} | \mathcal{X} = X(t)\right]$$

The optimal classification rule is to assign a new observation to that group that maximizes the posterior probability, i.e.

$$\hat{G}_X = \arg\max_{g \in \mathbb{G}} \hat{p}_g(X(t))$$

Table C.1 shows a list of the classification procedures described in this appendix and implemented in R. We have developed these classification methods that can be grouped in the following three subsections.

C.2.1 Classifiers based on regression models

All the regression models shown throughout this section that can deal with binary response could also lead to a classification rule with the same strategy. The core of the methods is to consider the problem of classification as a logistic (linear or additive) regression problem.

Let $\pi_{i,g}$ the expectation of G given $X_i(t)$ that will be modelled as:

$$\pi_{i,g} = \mathbf{P}[G = g | \mathcal{X} = X_i] = \frac{\exp\{\alpha_g + \int_T X_i(t)\beta_g(t)dt\}}{1 + \exp\{\alpha_g + \int_T X_i(t)\beta_g(t)dt\}} \quad , i = 1, \ldots, n$$

We implement the following classification models based on functional generalized regression models:

i. Functional generalized linear models (FGLM): `classif.glm()`.

ii. Functional generalized spectral additive models (FGSAM): `classif.gsam()`.

iii. Functional generalized kernel additive models (FGKAM): `classif.gkam()`.

These methods allow binary and multiclass classification through maximum probability.

Going back to the example of Section 2.3.3, the aim is to classify the fat content (smaller or larger than 15%, `Fat.bin`$= \mathbb{1}_{\{\text{Fat}>0.15\}}$) according to the second derivative of spectrometric curves `X.d2`.

```
data(tecator)
Fat.bin <- ifelse(tecator[["y"]][,"Fat"] > 15, 1, 0)
dataf <- data.frame(tecator[["y"]], Fat.bin)
names(dataf)[4] <- "Fat.bin"
X.d2 <- fdata.deriv(absorp, nderiv = 2)
ldata <- list("df" = dataf, "X.d2" = X.d2)
```

For generalized models, the arguments of the classification functions are analogous to those of the regression functions: `fregre.glm()`, `fregre.gsam()` and `fregre.gkam()`, that is, by means of an argument `formula` (describing the model), the argument `data` (containing the variables named in model) and a series of tuning parameters (see help of the function).

```
classif.glm(Fat.bin ~ X.d2, data = ldata)
##
## -Call:
## classif.glm(formula = Fat.bin ~ X.d2, data = ldata)
##
## -Probability of correct classification:  0.9767
classif.gsam(Fat.bin ~ s(X.d2), data = ldata)
##
## -Call:
## classif.gsam(formula = Fat.bin ~ s(X.d2), data = ldata)
##
## -Probability of correct classification:  1
classif.gkam(Fat.bin ~ X.d2, data = ldata)
##
## -Call:
## classif.gkam(formula = Fat.bin ~ X.d2, data = ldata)
##
## -Probability of correct classification:  0.9907
```

C.2.2 Non–linear classifiers

The `classif.kernel()` function uses the non–linear regression presented in Section 2.3.4 to compute the posterior probabilities $\hat{p}_g(X)$ for each group based on Kernel estimator.

$$\hat{p}_{g,h}(X(t)) = \frac{\mathbb{1}_{\{G_i=g\}}\sum_{i=1}^{n} K(h^{-1}d(X,\mathcal{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(X,\mathcal{X}_i))}$$

where $K(\cdot)$ is a kernel function, h is the smoothing parameter, and $d(\cdot,\cdot)$ is a metric or a semi-metric. The kernel is applied to a metric or semi-metric that provides non-negative values, so it is common to use asymmetric kernels. This estimator fulfills the following properties:

$$0 \leq \hat{p}_{g,h}(X) \leq 1, \sum_{g\in\mathbb{G}} \hat{p}_{g,h}(X) = 1$$

An alternative to kernel approximation could be k-nearest neighbour in `classif.knn()` function. The call for both classifiers is restricted to a single functional covariate like the corresponding regression model (`fregre.np()`) and now `group` is the argument of the response instead of `y`.

```
classif.knn(y = Fat.bin, group = X.d2)
##
## -Call:
## classif.np(group = group, fdataobj = fdataobj, h = knn,
## Ker = Ker.unif, metric = metric, type.CV = type.CV, type.S = S.KNN,
## par.CV = par.CV, par.S = par.S)
##
## -Optimal bandwidth: h.opt= 31 with highest probability of
##    correct classification: max.prob= 0.9767442
```

```
classif.kernel(y = Fat.bin, group = X.d2)
##
## -Call:
## classif.np(group = group, fdataobj = fdataobj, h = h, Ker = Ker,
## metric = metric, type.CV = type.CV, type.S = S.NW,
## par.CV = par.CV, par.S = par.S)
##
## -Optimal bandwidth: h.opt= 0.0004845957 with highest probability of
##    correct classification: max.prob= 0.9860465
```

C.2.3 DD^G-Classifier

The DD^G-Classifier fits nonparametric classification based on DD–plot (depth-versus-depth plot) for G dimensions ($G = g \times p$, g levels and p data depth). The main arguments of `classif.DD()` are:

i. `group`. Factor of length n with g levels.

ii. `fdataobj`. data.frame, fdata or list with the multivariate, functional or both covariates respectively.

iii. `depth`. Character vector specifying the type of depth functions to use. Table C.1 shows a list of all the implemented depth functions that are shown below according to the type of input data.

   (a) Type of depth function from functional data. If `depth= "mode"`, the function calls `depth.mode()` function where h–modal depth is computed, these options correspond to below examples.

   (b) Type of depth function from multivariate functional data. If `depth= "mode"`, the function calls `depth.modep()` function where h–modal depth is computed using a p–dimensional metric.

   (c) Type of depth function from multivariate data, see `Depth.Multivariate` help in R. If `depth="LD"`, the function calls `mdepth.LD()` function where the Likelihood depth is computed.

   The user can specify the parameters for depth function in `par.depth` argument.

iv. `classif`. Character vector specifying the type of classifier method to use, see the below options:

   (a) `"DD1"`, `"DD1"` and `"DD3"`: Search for the best separating polynomial of degree 1, 2 and 3 respectively.

   (b) `"MaxD"`: Maximum depth. Please note that the maximum depth classifier can be considered as a particular case of `"DD1"`, fixing the slope with a value of 1 (`par.classif= list(pol = 1)`).

   (c) `"glm"` and `"gam"`: Logistic regression is computed using Generalized Linear Models `classif.glm()` and Generalized Additive Models `classif.gsam()` respectively.

   (d) `"lda"` and `"qda"`: Linear or Quadratic Discriminant Analysis is computed using `lda()` or `qda()` respectively.

   (e) `"np"` and `"knn"`: Non-parametric Kernel and k-Nearest Neighbour classifier is computed using `classif.np()` and `classif.knn()` respectively.

The user can specify the parameters for classification method in `par.classif` argument.

The call to the DD–classifier is quite intuitive. Below are three examples with the same depth (argument `depth="mode"`) and different classifiers (argument `classif="glm"`, `classif="gam"` and `classif="np"` respectively).

```
classif.DD(group = Fat.bin, X.d2, depth = "mode", classif = "glm")
##
## -Call:
## classif.DD(group = Fat.bin, fdataobj = X.d2, depth = "mode",
## classif = "glm")
##
## -Probability of correct classification:  0.9721
```

```
classif.DD(group = Fat.bin, X.d2, depth = "mode", classif = "gam")
##
## -Call:
## classif.DD(group = Fat.bin, fdataobj = X.d2, depth = "mode",
## classif = "gam")
##
## -Probability of correct classification:  0.9767
```

```
classif.DD(group = Fat.bin, X.d2, depth = "mode", classif = "np")
##
## -Call:
## classif.DD(group = Fat.bin, fdataobj = X.d2, depth = "mode",
## classif = "np")
##
## -Probability of correct classification:  0.9721
```

Appendix D

Supplement to Chapter 5

This supplement contains some of the main contributions developed from the publication of paper Febrero–Bande and Oviedo de la Fuente (2012) (see Appendix D.1) and a R example of its use for functional regression with impact point selection (see Appendix D.2).

D.1 Functions added to *fda.usc* package

R Function	Impact Points Selection
LMDC.select()	Selects impact points of $X(t)$ using local maxima distance correlation
LMDC.regre()	Multivariate regression using the selected impact points

Table D.1: R functions for impact points selection

D.2 Impact point selection: R example

In the literature, the tecator data set has also been used to select the optimal points of the spectrometric curves, for example, in the NOVAS procedure proposed by Ferraty *et al.* (2010) and in wavelet-based LASSO approach by Zhao *et al.* (2012). In below example, we apply our LMDC approach for selected impact points of the second derivative of spectrometric curves. First, we create the input objects:

```
data(tecator)
y <- tecator[["y"]][["Fat"]]
X.d2 <- fdata.deriv(tecator[["absorp.fdata"]],nderiv =2,method ="fmm")
colnames(X.d2[["data"]]) <- paste0("X",round(X.d2[["argvals"]]))
df <- data.frame("y" = y, X.d2[["data"]])
```

Below, we apply the first part of LMDC procedure described in Chapter 5.2.2:

```
dc.raw <- LMDC.select("y", data = df, tol = 0.05, pvalue = 0.05,
plot=F, smo=T)
```

In the above code, the following steps have been taken:

1. Calculate the distance correlation $\mathcal{R}(t) = \{\mathcal{R}(X(t_j),Y)\}_{j=1}^{N}$, using the expression in (5.11). The related arguments are the following: `y`, name of the response variable; `covar`, vector with the names of the covariables (or points of impact), if it is missing, the function uses the names of `data` argument; and `data`, a data frame containing the scalar response and the potential N covaviables in the model.

2. With argument `smo=TRUE`, the smoothed distance correlation function $\{\hat{R}(t_j)\}_{j=1}^{N=100}$ is computed in order to avoid non relevant local maxima.

3. The predictive point t_j is selected if $\hat{R}(t_j)$ overtakes the value of threshold and the distance correlation t-test is significant (arguments `tol` and `pvalue` respectively).

4. The function orders the t_j's from highest to lowest values of $\mathcal{R}(\tilde{t}_j)$, that is $\hat{\mathcal{R}}(\tilde{t}_1) \geq \hat{\mathcal{R}}(\tilde{t}_2) > \ldots \geq \hat{\mathcal{R}}(\tilde{t}_{\tilde{N}})$, with $\tilde{N} = 13$.

```
# Preselected impact points
covar<-names(df)[-1][dc.raw[["maxLocal"]]]
covar

##  [1] "X933"  "X850"  "X1048" "X911"  "X951"  "X878"  "X886"
##  [8] "X1008" "X1022" "X1028" "X862"  "X995"  "X971"

length(covar)

## [1] 13
```

An optional step, we may verify whether the relationship between the response and the predictor variables is linear. To this aim, we apply a test of linearity that uses the Projected Cramer-von Mises statistic proposed by García-Portugués *et al.* (2014).

```
ftest <- flm.test(df[,-1], df[,"y"], verbose = F, plot.it = F)
ftest
##
##  PCvM test for the functional linear model using optimal PLS basis
##  representation
##
## data:  Y=<X,b>+e
## PCvM statistic = 193.67, p-value < 2.2e-16
```

As the `p-value<0.05` it is recommended to adjust a non-linear model. Then, the below code fits a non-linear model to the response of interest Y using the vector of covariates $\mathbf{Z} = \{X(\tilde{t}_1), \ldots, X(\tilde{t}_{\tilde{N}=13})\}$. To carry it out the `LMDC.regre()` function is used. The user must specify the type model has been selected in `method` argument. We propose to apply a forward stepwise additive regression method `method="gam"` to determine the significant covariates, taking advantage of the fact that the local maxima have been ordered.

```
if (ftest[["p.value"]] > 0.05) {
# Linear relationship, step-wise lm is recommended
out <- LMDC.regre("y", covar, df, pvalue = 0.05, method = "lm",
 plot = F, verbose = F)
}  else {
# Non-Linear relationship, step-wise gam is recommended
# Arguments for gam model
par.method<-list("k" = -1, method = "GCV.Cp")
out <- LMDC.regre("y", covar, df,  pvalue= 0.05, method = "gam",
 plot = F, verbose = F,  par.method=par.method)   }
```

As a result, the final covariates $\mathbf{Z} = \{X(t_{933}), X(t_{1048}), X(t_{911}), X(t_{951}), X(t_{1008}), X(t_{1022}), X(t_{995}), X(t_{971})\}$ with $\tilde{\tilde{N}} = 8$ fulfills $\tilde{\tilde{N}} \leq \tilde{N} \leq N$,

```
print(impact.points <- out[["xvar"]])
## [1] "X933"  "X1048" "X911"  "X951"  "X1008" "X1022" "X995"  "X971"
length(impact.points)
## [1] 8
```

and the fitted model is:

```
#
summary(out[["model"]])
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(X933) + s(X1048) + s(X911) + s(X951) + s(X1008) + s(X1022)
##  + s(X995) + s(X971)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.1423     0.0439   413.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df       F  p-value
## s(X933)   7.579  8.298 153.435  < 2e-16 ***
## s(X1048)  7.112  7.966   2.715  0.00674 **
## s(X911)   8.675  8.935  14.973  < 2e-16 ***
## s(X951)   6.060  7.204  21.895  < 2e-16 ***
## s(X1008)  7.247  7.963   4.469 8.53e-05 ***
## s(X1022)  4.655  5.777   7.866 3.39e-07 ***
## s(X995)   8.372  8.754   5.307 3.11e-06 ***
## s(X971)   7.773  8.260   2.837  0.00669 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.997   Deviance explained = 99.8%
## GCV = 0.56917  Scale est. = 0.41437   n = 215
```

Finally, we plot in Figure D.1 the impact points selected (in red) in the second derivative of spectrometric curves.

```
plot(X.d2, col = 1)
abline(v=substr(impact.points, 2, 6), col = 2, lty = 2)
```

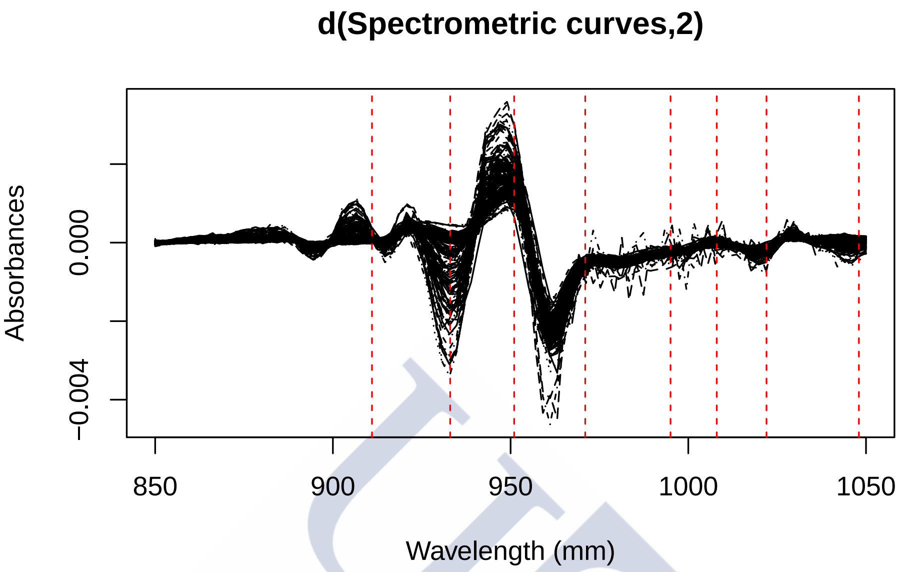

Figure D.1: In black, spectrometric curves (2nd derivative) and in red, impact points selected.

References

Cuesta-Albertos, J. A., Febrero-Bande, M., and Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *TEST*, 26(1):119–142

Febrero–Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package `fda.usc`. *J Stat Softw*. 51(4):1–28.

Febrero–Bande, M., González–Manteiga, W. (2013). Generalized additive models for functional data. *TEST*, **22**(2),278-292.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer-Velag, New York. Theory and practice.

Ferraty, F., Hall, P., and Vieu, P. (2010). Most predictive design points for functional data predictors. *Biometrika*, 94(4):807–824.

García-Portugués, E., González-Manteiga, W., and Febrero-Bande, M. (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23(3):761–778.

Müller, H.G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 426–437.

Ordóñez, C., Oviedo de la Fuente, M., Roca-Pardiñas, J., and Rodríguez-Pérez, J. R. (2018). Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. *Chemometrics and Intelligent Laboratory Systems*, 173:41–50.

Oviedo de la Fuente, M., Febrero-Bande, M., Muñoz, M. P., and Domínguez, À. (2018). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PloS one*, 13(4):e0194250.

Ramsay J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer Series in Statistics, second edition. Springer-Velag, New York.

Wood, S.N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC.

Zhao, Y. and Ogden, R.T. (2015). Wavelet-based weighted lasso and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics*, 24(3):655–675.

Zhao, Y., Ogden, R.T., and Reiss, P.T. (2012). Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

Resumen extendido en castellano

En los últimos años la estadística y la informática aplicada a diversos campos ha producido importantes cambios tecnológicos; los cuales consisten en equipos dotados de mayor rapidez que consecuentemente proporcionan mediciones más precisas y rápidas. Esta evolución tecnológica ha modificado algunos de los paradigmas en los que se basa la estadística clásica, por ejemplo, aquellos en los que el número de observaciones en un conjunto de datos es mayor que el número de variables. Además, en muchas áreas se ha comenzado a trabajar con grandes bases de datos en las cuales es cada vez más común registrar las observaciones de una variable aleatoria en un intervalo continuo. Por ejemplo, en campos como la espectroscopia, el resultado de la medición es una curva que se ha evaluado en al menos un centenar de puntos. Este tipo de datos, generalmente llamados datos funcionales, surgen naturalmente en muchas disciplinas, en este sentido, podríamos hablar de curvas de precios de acciones intradía en energía, curvas de demanda de electricidad en el medio ambiente, la evolución de la tasa de gripe estacional en epidemiología, curvas de reflectancia de la hoja de vino en la agricultura, etc.

Esta tesis tiene como punto de partida las contribuciones básicas en la literatura sobre datos funcionales correspondientes a los libros de Ramsay and Silverman (2005) y Ferraty and Vieu (2006), respectivamente. Sin embargo, la novedad del análisis de datos funcionales (ADF) abre una amplia gama de posibles líneas de investigación para hacer frente a una complejidad en los datos cada vez mayor. Las revisiones recientes de Cuevas (2014) y Ferraty et al. (2011) ofrecen una perspectiva muy amplia sobre el estado del arte en el ADF.

El primer capítulo de esta tesis introduce al lector sobre el estado del arte en el ADF y resume el contenido de la tesis. Este documento se trata de una tesis por compendio de publicaciones por lo que cada capítulo contiene un artículo original con su propio resumen, secciones, apéndices y referencias. En particular, los artículos correspondientes a los Capítulos 2-6 se han publicado en revistas estadísticas del JCR con un factor de alto impacto. El Capítulo 7 contiene las conclusiones de la tesis y finalmente los Apéndices A, B, C y D recogen el material suplementario, como las funciones desarrolladas en R con un ejemplo de su uso. Por supuesto, todos los avances metodológicos se han presentado en conferencias nacionales e internacionales en el campo del análisis de datos funcionales.

A continuación se muestran tanto los objetivos generales y específicos de la tesis, seguidamente se muestran el resumen de cada capítulo (artículo) y finalmente, se describen las conclusiones y trabajo futuro derivados de la elaboración de la tesis.

Objetivos

Los objetivos de la tesis recogidos en el plan de investigación del Programa de Doctorado en Estadística e Investigación Operativa de la Universidade de Santiago de Compostela son los siguientes:

OBJ1 Revisión del estado del arte actual en el ADF e implementación de los procedimientos más relevantes en R como es la representación en base fija (B-splines, Fourier, Wavelet) o basada en los datos (componentes principales PC, PLS), suavización tipo kernel, medidas de correlación entre objetos (correlación de distancias), detección de datos atípicos en datos funcionales multivariantes.

OBJ2 Desarrollo de un marco general para extender los modelos FGLM al caso en el que la respuesta sea un recuento o una curva.

OBJ3 Extender los modelos FGLM para analizar datos funcionales que presenten una dependencia, espacial o temporal. En esta tesis se estudian los modelos LM y GLS con el fin de generalizarlos mediante el modelo GLS funcional. Además se propone una versión iterativa que permite la estimación de estructuras de dependencia en los errores.

OBJ4 Implementación de los modelos de clasificación supervisada, como la regresión logística (lineal o aditiva) y el clasificador *k*NN funcional, entre otros. Los avances metodológicos se centran en los métodos basados en medidas de profundidad como es el caso del clasificador DD^G que extiende el DD-plot (Li et al., 2012) y permite la clasificación de datos complejos como imágenes hiperespectrales y datos funcionales multivariantes.

OBJ5 Desarrollo de algoritmos que permiten la selección de covariables en contextos de regresión funcional. Para ello se utiliza la correlación de distancia propuesta por (Székely et al., 2007).

Capítulo 2. Nuevas contribuciones para el Análisis de Datos Funcionales usando la libreria fda.usc

Este capítulo es una revisión detallada del estado del arte relacionado con el ADF y los métodos estadísticos que pueden aplicarse a un conjunto de datos funcionales. El núcleo de los principales avances de este capítulo consistió en desarrollar un marco común en R (en forma de paquete) que aglutinaba, de forma estructurada y homogénea, las propuestas de las dos tendencias principales en el análisis de datos funcionales y principales referencias bibliográficas; el libro de Ramsay and Silverman (2005) desde una perspectiva paramétrica y el libro de Ferraty and Vieu (2006) desde una no-paramétrica. Además, se ha realizado una labor constante de mantener, incorporar e integrar (siguiendo los estándares del lenguaje de programación R) los procedimientos más relevantes publicados en la literatura de ADF con nuestros propios desarrollos. Todo ello en una interfaz simple, común, operativa y de fácil acceso, como es el paquete de R `fda.usc`, Febrero–Bande and Oviedo de la Fuente (2012).

Algunos de estos procedimientos son:

i. Representación en una base (fija o basada en los datos) de un conjunto de datos funcionales (DF).

ii. Suavizado de un DF mediante una función kernel o una base fija.

iii. Calculo de métricas o semimétricas entre DF univariantes.

iv. Calculo de medidas de profundidad para un conjunto de DF.

v. Procedimientos de detección de datos funcionales atípicos.

La segunda parte del capítulo tratan los modelos de regresión funcional, que son aquellos en el que al menos una de las variables (ya sea una variable de predicción o una variable de respuesta) es funcional. Una de las principales contribuciones de esta tesis es la implementación de varios modelos de regresión funcional (FRM) en los que la respuesta o al menos una de las covariables tiene una naturaleza funcional. Los FRM se han estudiado ampliamente en la segunda parte del Capítulo 2, especialmente cuando la respuesta es escalar y las covariables son funcionales.

Partimos pues de un enfoque general a partir del cual se deducirán los diferentes modelos a medida que aumenta su complejidad.

Sea $Y \in \mathbb{R}$ la respuesta escalar, $\mathbf{Z} \in \mathbb{R}^p$, el vector de p covariables y $\mathcal{X} = \{X^j(t)\}_{j=1}^q$ las q covariables funcionales el modelo general planteado es:

$$Y = m(\mathbf{Z}, \mathcal{X}) + \varepsilon, \tag{7.1}$$

donde $m(\cdot)$ es una función desconocida que combina las covariables con la respuesta y ε es un proceso de error.

Durante la tesis se relaciona el modelo anterior con algunos modelos importantes introducidos en la literatura en los últimos años. Así pues, la primera propuesta asume la linealidad en la esperanza condicional de la respuesta y las covariables. Este es un modelo lineal funcional (FLM), que para el caso de la respuesta escalar Y y una sola covariable funcional $\mathcal{X} = X(t)$ se puede escribir como: $\mathbf{E}[Y|\mathcal{X}] = \alpha + \langle \beta, \mathcal{X} \rangle$, donde α es el parámetro de intersección y $\beta = \beta(t)$ es el parámetro funcional, ambos a estimar. En el FLM la idea principal es la proyección de cada par $X(t)$ y $\beta(t)$ en un número finito de elementos de una base funcional que pueden elegirse fijos de antemano, como B-spline, Fourier o Wavelet (ver por ejemplo los trabajos de Cardot et al. (2003) y Ramsay and Silverman (2005)) o en función de datos, como en el caso de las componentes principales (PC) funcionales o mínimos cuadrados parciales (PLS) funcionales, (ver Cardot et al. (1999)y Aguilera et al. (2010) respectivamente).

El modelo dado por la ecuación (7.1) también puede involucrar predictores funcionales, pero, de una manera flexible asumiendo una relación no-lineal entre los predictores y la respuesta. Al igual que en el contexto de la regresión estándar con covariables escalares, se pueden obtener ajustes más precisos modelando dicha relación no lineal. En particular, el modelo de regresión no paramétrico con una única covariable funcional ha sido ampliamente estudiado en la literatura, como se puede ver en el libro de Ferraty and Vieu (2006) y artículos relacionados. La estimación de $m(\cdot)$ se da usando diferentes procedimientos basados en técnicas de suavizado como splines y suavizado de kernel (Nadaraya–Watson, Lineal Local y kNN). En el Capítulo 2, también se trata el modelo de regresión funcional parcialmente lineal (FPLM) introducido y estudiado en Aneiros-Pérez and Vieu (2006). El modelo FPLM generaliza el modelo no paramétrico con una sola covariable funcional $\mathcal{X}$ permitiendo incorporar variables escalares exógenas $\mathbf{Z} \in \mathbb{R}^p$ mediante una componente lineal.

Estos desarrollos se refieren al artículo titulado "**Statistical computing in functional data analysis: the R package *fda.usc***" (Febrero–Bande and Oviedo de la Fuente, 2012) publicado en la revista *Journal of Statistical Software* en 2012 para el cual la información asociada (incluyendo artículo y la versión 1.0 del paquete *fda.usc*) está disponible en el DOI: http://dx.doi.org/10.18637/jss.v051.i04, vea Febrero–Bande and Oviedo de la Fuente (2012).

En varias aplicaciones, el modelo lineal funcional (FLM) puede ser demasiado restrictivo, por ejemplo cuando la respuesta es binaria o un recuento. Una extensión natural de los modelos FLM es el modelo lineal generalizado funcional (FGLM) propuesto por Müller and Stadtmüller (2005). En el marco de los modelos generalizados se asume generalmente que la respuesta escalar puede elegirse dentro del conjunto de distribuciones pertenecientes a la familia exponencial. Este enfoque general incluye, entre otros, el caso del FLM, la regresión funcional de Poisson y la regresión binomial funcional. En relación a la regresión binomial, también conocida a regresión logística, puede utilizarse como un procedimiento de clasificación de datos funcionales.

Los Apéndices A y C recogen el uso en R del modelo FGLM y también dos interesantes extensiones que permiten estimaciones no-lineales de los efectos de las covariables en la respuesta: La primera corresponde con el modelo FGSAM propuesto por Müller and Yao (2008) ya que extiende los modelos GAM al caso funcional mediante la representación en una base del dato funcional. En la práctica, el procedimiento es bastante flexible porque sigue la estructura de la función `gam()` del paquete `mgcv`. En la segunda, el modelo aditivo kernel generalizado funcional (FGKAM) propuesto por Febrero-Bande and González-Manteiga (2013) ajusta las funciones $m_j(\cdot)$ mediante estimaciones no-parámetricas tipo kernel siguiendo las pautas del algoritmo backfiting. Tanto los procedimientos FGSAM como FGKAM se han desarrollados en el paquete *fda.usc*.

Otros procedimientos relevantes, como el modelo de respuesta funcional tratado por Chiou et al. (2004) y herramientas tales como la identificación de observaciones influyentes también se han implementado en R.

Capítulo 3. Modelo de regresión funcional con errores dependientes

El objetivo de este capítulo es ampliar las ideas de los métodos de mínimos cuadrados generalizados (GLS, en sus siglas en inglés) a los modelos de regresión funcional (FRM) con respuesta escalar. En un marco multivariante, los estimadores de GLS nos permiten incorporar una amplia lista de estructuras de covarianza para el término de error en modelos de regresión. Como ejemplo, estos modelos pueden incluir correlación temporal, dependencia espacial, heterocedasticidad o incluso, efectos aleatorios. En esta tesis, la metodología GLS se extiende al modelo de regresión funcional, especialmente al caso en que la estructura de covarianza es desconocida y debe estimarse. También se propone una versión iterativa del estimador GLS (llamado iGLS) que puede ayudar a modelar estructuras de dependencia complicadas.

Los estudios de simulación muestran que los estimadores GLS proporcionan mejores estimaciones de los parámetros asociados al modelo de regresión que con los modelos clásicos, obtienen resultados extremadamente buenos desde el punto de vista predictivo y son competitivos con el enfoque clásico de series de tiempo. Esto se ha corroborado en el ejemplo real de predicción de la incidencia de la gripe en Galicia (España) utilizando la información meteorológica. Así pues, los modelos GLS han resultado útiles cuando la historia reciente de la incidencia de la gripe no está disponible (por ejemplo, por retrasos en la comunicación con informantes de salud) y la predicción debe construirse corrigiendo la dependencia temporal de los residuos y utilizando variables más accesibles. Además, para construir el modelo, se utilizó la medida de correlación de distancia $\mathcal{R}$ propuesta por Székely et al. (2007) para seleccionar la información relevante (variables multivariantes y funcionales) utilizadas en la predicción de la tasa de gripe.

El artículo titulado "**Predicting seasonal influenza transmission using functional regression models with temporal dependence**" publicado en *PLoS ONE* en 2018 incluye estos desarrollos; su información asociada (incluido el artículo y la versión 1.4 del paquete `fda.usc`) está disponible en el DOI: `https://doi.org/10.1371/journal.pone.0194250`, ver Oviedo de la Fuente et al. (2018).

Este tipo de modelos son extremadamente útiles para los administradores de salud en la asignación de recursos de antemano para gestionar epidemias. Desde 2016 este procedimiento se está aplicando a la predicción de la tasa de gripe en Cataluña (Basile et al., 2018).

Capítulo 4. Técnicas de clasificación basadas en las medidas de profundidad funcional

El Capítulo 4 está dedicado principalmente a métodos de clasificación supervisados que utilizan clasificadores basados en la información derivada de la profundidad de los datos funcionales. Este capítulo tiene como objetivo ampliar el clasificador DD-plot propuesto por Li et al. (2012) de tres maneras:

i. El procedimiento original se basa en la representación gráfica del par $(D_P(x), D_Q(x)) \in \mathbb{R}^2$ para cada $x \in \mathbb{R}^p$. Este par no es más que el vector 2-dimensional de la profundidad de x respecto a dos medidas de probabilidad **P** y **Q** en $\mathbb{R}^p$ o dos clases/grupos (independientemente de la dimensión p). Se extiende el DD–plot al permitir la clasificación en $G \geq 2$ clases o grupos.

ii. El DD-plot utiliza un clasificador en polinomios de grado 1, 2 o 3. Este clasificador tiene algunas limitaciones:

- la cantidad de modelos polinomiales para crear la regla de clasificación es $2\binom{G}{2}\binom{N}{k}$ que puede ser extremadamente elevada, donde G es el número de clases, k el grado del polinomio y N el tamaño de la muestra.

- los polinomios siempre se calculan en fronteras entre los grupos que no permiten la construcción de zonas que diferencien la asignación a un grupo u otro.

El clasificador DD^G ofrece una solución unificada y en general con menor coste computacional a estos inconvenientes al aplicar métodos de clasificación regulares (como *k*NN, clasificadores lineales o cuadráticos, particiones recursivas,...). Además, algunos de estos clasificadores son particularmente útiles ya que proporcionan información basada en el diagnóstico del modelo.

iii. El clasificador DD^G integra de forma unificada diversas fuentes de información (vectores con la profundidades de los datos, datos funcionales multivariantes,...) en el procedimiento de clasificación.

Para llevar a cabo estas mejoras ha sido necesaria una revisión mejorada de varias profundidades de datos funcionales que va acompañada de un amplio estudio de simulación y de aplicaciones en los conjuntos de datos más utilizados en la literatura.

Este Capítulo hace referencia al artículo titulado "**The DD^G-classifier in the functional setting**" publicado en la revista *TEST* en 2017, disponible en: `https://link.springer.com/article/10.1007%2Fs11749-016-0502-6`; su material complementario está disponible en la versión en línea en el DOI: `https://doi.org/10.1007/s11749-016-0502-6`, vea Cuesta-Albertos et al. (2017).

Capítulo 5. Regresión funcional con puntos de impacto

En este trabajo estudiamos la utilidad de la correlación de distancia (Székely et al., 2007) como un método para seleccionar puntos de impacto de un predictor funcional. Ello no requiere suponer a priori de un modelo de regresión; sino que simplemente buscamos los máximos locales de la función de correlación de distancia. El trabajo está estructurado de la siguiente manera: en primer lugar, proporcionamos un breve resumen del modelo de regresión funcional paramétrica (Ramsay and Silverman, 2005) y no-paramétrica (Ferraty and Vieu, 2006), en segundo lugar, proporcionamos una breve explicación de los tres métodos utilizados para determinar las longitudes de onda óptimas. El primero se basa en técnicas de suavizado tipo kernel (Ferraty and Romain, 2010) para seleccionar los puntos de diseño más predictivos. El segundo es un procedimiento basado en la representación en una base wavelet y la aplicación del procedimiento LASSO a los coeficientes de la base (Zhao et al., 2015), y último, que corresponde a nuestra propuesta (Ordóñez et al., 2018), esta basado en el cálculo de máximos locales en la función de correlación de distancias. En el estudio hemos aplicamos los métodos considerados a datos simulados y a un conjunto de datos real.

Este capítulo hace referencia al artículo titulado "**Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach.**" publicado en la revista *Chemometrics and Intelligent Laboratory Systems.* en 2017, disponible en: `https://www.sciencedirect.com/science/article/pii/S0169743917304768?via%3Dihub`, cuyo material complementario está disponible en la versión en línea en el DOI: `https://doi.org/10.1016/j.chemolab.2017.12.001`, vea Ordóñez et al. (2018).

Capítulo 6. Selección de variables en modelos de regresión y clasificación funcional

Este capítulo considera el problema de la selección de variables en modelos de regresión en el caso de diferentes tipos de predictores (escalares, multivariantes, funcionales, direccionales, etc.). Nuestra propuesta comienza con un modelo nulo y selecciona de manera secuencial si una nueva variable se incorpora al modelo en función de la correlación de distancia que tenga con el residuo del modelo. Para datos multivariantes, hemos comparado nuestra propuesta en los mismos escenarios de Yenigün and Rizzo (2015) y en un escenario mixto con variables funcionales y escalares. Además, el último ejemplo de resultados numéricos está relacionado con un problema de clasificación; en la que la respuesta es binomial. El procedimiento se aplicó a un problema real relacionado con el Mercado Energético Ibérico (Precio y Demanda) donde el número de posibles covariables es realmente grande.

Este capítulo hace referencia al artículo titulado "**Variable selection in Functional Additive Regression Models.**", vea (Febrero-Bande *et al.* (2018)), publicado en Computational Statistics, DOI: `https://doi.org/10.1007/s00180-018-0844-5`. Este trabajo proporciona novedades respecto del capítulo de libro (Febrero-Bande *et al.*, 2017, pp. 113-122).

Conclusiones y trabajo futuro

A continuación se resumen la conclusiones de la tesis, muchas de ellas ya estaban recogidas previamente en cada uno de los capítulos.

i. La primera tarea ha sido construir un entorno de software donde integrar todas las herramientas de ADF (análisis exploratorio, regresión, clasificación supervisada y no supervisada, analisis de la varianza,...). Para ello se ha creado el paquete de R `fda.usc` que aglutina dichos procedimientos. En particular, el Capítulo 2 resume los procedimiento programados en R de forma eficiente y general para que los usuarios puedan extender y reproducir en el futuro con sus propios conjuntos de datos. A lo largo de la tesis, se han ido incorporando en el paquete `fda.usc` las nuevas propuestas implementadas como el cálculo de distancias y las medidas de profundidad para conjuntos de datos funcionales multivariantes, tal y como se recoge en el Capítulo 4. En esa linea, la detección de valores atípicos en datos complejos o en al menos en datos funcionales multivariantes es otro tópico donde es preciso investigar nuevas propuestas. El trabajo propuesto por Hubert et al. (2015) puede ser un buen punto de partida para avanzar en los desarrollos teóricos que puedan ser viables en la práctica (por ejemplo, que puedan implementarse en R).

   Esta tesis considera el análisis de datos funcionales en el dominio dimensional 1D. La extensión a 2D puede suponer afrontar problemas similares, tales como: representación, regularización, visualización y suavizado.

ii. En segundo lugar, se ha desarrollado un marco general para extender los modelos de regresión funcional para el caso de una respuesta de la familia exponencial. El Apéndice A amplía los modelos GLM y GAM al caso funcional. Las funciones se han incluido en el `fda.usc` y se han presentado en varias comunicaciones a congresos.

iii. El Capítulo 3 propone extender el modelo lineal funcional para analizar datos funcionales con dependencia espacial o temporal. Así pues se ha propuesto un modelo GLS para covariables funcionales así como una versión iterativa (llamada iGLS) que puede ser de ayuda para modelar estructuras de correlación complejas. Como trabajo futuro, se planea extender las estructuras de dependencia a problemas más complejos como la dependencia espacio-temporal. Además, el tratamiento de series temporales funcionales está todavía sin abordar desde el punto de vista de software, una posibilidad es seguir en la linea de investigación relacionada con los modelos de regresión funcional con respuesta funcional.

iv. El Capítulo 4 recoge una nueva propuesta de clasificación basada en profundidades que además ha sido comparada con las técnicas clásicas de clasificación de datos funcionales (que también han sido incluidas en *`fda.usc`*).

   Actualmente, se está trabajando en adaptar los procedimientos disponibles al problema clasificación: de grupos con diferente tamaños de muestra, de datos funcionales multivariados y de datos complejos (como el caso de las imágenes hiperespectrales).

v. El Capítulo 5 propone un algoritmo que permite la selección de covariables (puntos de impacto) en contextos de regresión con una covariable funcional. Su rendimiento muestra que es una buena alternativa al trabajo de Ferraty *et al.* (2010). La principal ventaja de nuestro enfoque es que es una regla incremental, en la que lista de covariables candidatas está ordenada por los máximos locales de la correlación de distancia entre los puntos de impacto y la respuesta.

vi. El Capítulo 6 propone un marco general para el caso de la selección de variables en regresión con información funcional, escalar o categórica. Hasta donde tenemos conocimiento, no hay trabajos previos que proporcionen un enfoque general como el que hemos desarrollado. Nos hemos limitado a modelos aditivos ya que ofrecen un compromiso equilibrado entre la capacidad de predicción y la simplicidad incluso para un gran número de covariables.

    Aunque hay una versión beta de esta propuesta, se esta trabajando en la implementación final del algoritmo (versión estable) que será próximamente accesible a la comunidad científica a través de la librería `fda.usc`.

References

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems,*, 104(2), 289-305.

Aneiros-Pérez, G. and Vieu, P. (2006). Semi-functional partial linear regression. *Statistics & Probability Letters*, 76(11):1102–1110.

Basile, L., Oviedo de la Fuente, M., Torner, N., Martínez, A., and Jané, M. (2018). Real-time predictive seasonal influenza model in catalonia, spain. *PloS one*, 13(3):e0193651.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–592.

Chiou, J.-M., Müller, H.-G., and Wang, J.-L. (2004). Functional response models. *Statistica Sinica*, pages 675–693.

Cuesta-Albertos, J. A., Febrero-Bande, M., and Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *Test*, 26(1):119–142.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, 47(1):111–122.

Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *Test*, 22(2):278–292.

Febrero-Bande, M., Gonzlez-Manteiga, W. and Oviedo de la Fuente M. (2017). Variable selection in Functional Additive Regression Models. *In Functional Statistics and Related Fields*, 113-122. Springer, Cham.

Febrero-Bande, M., González-Manteiga, W., and Oviedo de la Fuente M. (2018). Variable selection in Functional Additive Regression Models. *Computational Statistics*, DOI: `https://doi.org/10.1007/s00180-018-0844-5`.

Febrero–Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package `fda.usc`. *J Stat Softw*. 51(4):1–28.

Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2011). Recent advances on functional additive regression. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 97–102. Springer.

Ferraty, F. and Romain, Y. (2010). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press Oxford.

Ferraty, F., Hall, P., and Vieu, P. (2010). Most predictive design points for functional data predictors. *Biometrika*, 94(4):807–824.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). *DD*–classifier: Nonparametric classification procedure based on *dd*–plot. *Journal of the American Statistical Association*, 107(498):737–753.

Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805.

Müller, H. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Ordóñez, C., Oviedo de la Fuente, M., Roca-Pardiñas, J., and Rodríguez-Pérez, J. R. (2018). Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. *Chemometrics and Intelligent Laboratory Systems*, 173:41–50.

Oviedo de la Fuente, M., Febrero-Bande, M., Muñoz, M. P., and Domínguez, À. (2018). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PloS one*, 13(4):e0194250.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Yenigün, C. D. and Rizzo, M. L. (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, 85(8):1692–1705.

Zhao, Y., Chen, H., and Ogden, R. T. (2015). Wavelet-based weighted lasso and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics*, 24(3):655–675.

List of Figures

List of Tables