UNIVERSIDAD CARLOS III DE MADRID

Ph.D. Dissertation

The Mahalanobis distance for functional data with applications in statistical problems

Esdras Joseph

Advisors:

Pedro Galeano

Rosa E. Lillo

Department of Statistics

Leganés, March 2015

UNIVERSIDAD CARLOS III DE MADRID
DEPARTAMENTO DE ESTADÍSTICA
DOCTORADO EN INGENIERÍA MATEMÁTICA

TESIS DOCTORAL

The Mahalanobis distance for functional data with applications in statistical problems

Esdras Joseph

Directores:

Pedro Galeano

Rosa E. Lillo

Leganés, Marzo de 2015

Jurado asignado:

Presidente: .....

Secretario: .....

Vocal: .....

Directores:

Pedro Galeano	**Rosa E. Lillo**

This dissertation was written in the Department of Statistics at Universidad Carlos III de Madrid under the advise of the Professors Pedro Galeano and Rosa E. Lillo Rodríguez. The author was supported by a scholarship for master studies (UC3M) and was subsequently hired as a teaching and research assistant. Besides, the author and the advisors had the partial support of the following research project: Spanish Ministry of Economy and Competition grant ECO2012-38442.

A mi esposa

y

A mi mamá

Success is not final, failure is not fatal: it is the courage to continue that counts.

Winston Churchill

Motivation is what gets you started. Habit is what keeps you going.

Jim Rohn

Don't try to be original, just try to be good.

Paul Rand

Agradecimientos

Quisiera decir unas palabras a aquellas personas que de alguna forma u otra me han guiado y apoyado para sacar adelante esta tesis.

En primer lugar, mis más profundos y sinceros agradecimientos a mis directores de tesis Pedro Galeano y Rosa E. Lillo por la orientación, el seguimiento y la supervisión continúa de esta tesis. Quisiera dar las gracias a ellos sobre todo por la motivación y el apoyo recibido durante los años en los que hemos trabajado.

Debo agradececer de manera especial a Dios y a la iglesia Bautista Bethesda. A mi madre Jeannine Jean mil gracias por sus oraciones y consejos de seguir adelante ante cualquier dificultad. Dedico este trabajo a Marie Magdala mi esposa, por su apoyo y ánimo que me brinda diariamente para alcanzar nuevas metas, tanto como profesionales y personales. Estas dos mujeres sin duda son lo mejor y lo más importante de mi vida. Quiero extender un sincero agradecimiento a mis hermanos Marie Claude, Jonas, Joham y Francy por estar siempre a mi lado guiándome a pesar de la distancia.

Quisiera hacer extensiva mi gratitud al Departamento de Estadística por apoyarme financieramente para llevar a cabo este presente trabajo. No puedo olvidar en mis agradecimientos al equipo de profesores y al personal administrativo de dicho departamento: Gema, Susana y Paco por los servicios que me han brindado desde mi primer día en España. A mis compañeros de doctorado por los momentos que hemos compartido y que nunca podré olvidar; que sin ellos hubiese sido imposible llegar hasta donde estoy ahora: Gabi, Nicola, Mei, Joanna. Un especial agradeciemento a esas personas por todo su apoyo y consejo: Henry, Dalia, Lee y Willy.

Finalmente, quiero dar las gracias al profesor Mijail Borges-Quitana de la universidad de Oriente de Santiago de Cuba y a mis amigos Jean-Marie, Chrisner, Amazan, Cacoq, Junior, Calixte por sugerirme hacer una carrera como investigador.

A TODOS LOS QUE SE SIENTAN MENCIONADOS EN LA DEDICATORIA MIL GRACIAS.

Contents

List of Figures

Index of tables

The Mahalanobis distance for functional data with applications in statistical problems

Ph.D. Dissertation

Abstract

Esdras Joseph

Department of Statistics

Universidad Carlos III de Madrid

Functional data refer to data which consist of curves evaluated at a finite subset of some interval in the real line. In this thesis, we deal with this type of data, focusing on the notion of functional distance. In the literature, there is few references to the role played by distances between functional data. Recently, Ferraty and Vieu [20] have proposed some semi-metrics well adapted for sample functions. However, common distances frequently used for multivariate data analysis such as the Mahalanobis distance proposed by Mahalanobis [39], have not been extended to the functional framework. This issue motivated this thesis and its main contribution is to enlarge the number of available functional distances by introducing a new semi-distance that generalizes the usual Mahalanobis distance. The use of functional distances is important in many different problems, including supervised classification and hypothesis testing. Then the other contributions in this dissertation is to propose new procedures based on the combination of those methods with the functional Mahalanobis semi-distance as in the multivariate context.

The thesis is organized as follows. In Chapter 1 we review the formal definition of functional data as well as the notion of functional principal components which is an important tool for some of the concepts that will be seen in this dissertation. We also offer a brief historical summary of distances in the multivariate context and how the concept of distance has been extended to FDA. Finally, we recall some functional methods for which

the notion of distance can be very useful, e.g., supervised and unsupervised classification, hypothesis testing, prediction and the concept of density function for functional data.

In Chapter 2, we present a new semi-distance for functional observations that generalizes the Mahalanobis distance for multivariate datasets to the functional framework. We also shown the main characteristics of the functional Mahalanobis semi-distance. In order to illustrate the applicability of this measure of proximity between functional observations, we develop new versions of several well known functional classification procedures using the functional Mahalanobis semi-distance. We illustrate the performance of the new semi-distance with simulated and two real data examples indicating that the classification methods used in conjunction with the functional Mahalanobis semi-distance give better results than other well-known functional classification procedures.

In Chapter 3, we derive two-sample Hotelling's T^2 statistics for testing the equality of means in two samples independently drawn from two functional distributions. The statistics that we propose are based on the functional Mahalanobis semi-distance and, under certain conditions, their asymptotic distributions are chi-squared, regardless the distribution of the functional random samples. We provide the link between the two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance and statistics based on the functional principal components semi-distance. The behavior of all these statistics is analyzed by means of an extensive Monte Carlo study and the analysis of a real data set collected in climatology. The results appear to indicate that the two-sample Hotelling's T^2 statistics outperform in terms of power those based on the functional principal components semi-distance.

Finally, Chapter 4 is dedicated to some summary and some possible future research lines of the work presented in this thesis.

Distancia de Mahalanobis para datos funcionales con aplicaciones en problemas estadísticos

Tesis Doctoral

Resumen

Esdras Joseph

Departamento de Estadística

Universidad Carlos III de Madrid

El término de datos funcionales hace referencia a datos que en esencia son curvas, pero que están evaluadas en un subconjunto finito de algún intervalo de la recta real. Esta tesis trata sobre datos funcionales, centrándose en la noción de distancia funcional. En la literatura, las distancias entre datos funcionales no han sido muy tratadas. Recientemente, Ferraty y Vieu [20] han propuesto algunas semi-distancias adaptadas para muestras de funciones. Sin embargo, distancias comúnmente utilizadas para el análisis de datos multivariantes, tales como la distancia de Mahalanobis propuesta por Mahalanobis [39], no han sido extendidas al marco funcional. Esta tesis está motivada por esta cuestión y su principal contribución es ampliar el número de distancias funcionales disponibles introduciendo una nueva semi-distancia que generaliza la distancia de Mahalanobis. El uso de distancias funcionales es importante en algunos problemas estadísticos, incluyendo clasificación supervisada y contrastes para diferencias de medias. Las restantes contribuciones de esta tesis consisten en proponer nuevos procedimientos basados en la combinación de estos métodos con la semi-distancia de Mahalanobis funcional.

La tesis tiene la siguiente estructura. En el Capítulo 1 se revisa la definición formal de datos funcionales, así como la noción de componentes principales funcionales que es una herramienta importante para algunos de los conceptos desarrollados en los capítulos de contribución. Se ofrece también un breve resumen histórico de distancias en el contexto multivariante, y cómo el concepto de distancias ha sido extendido al análisis de datos funcionales. Finalmente, se recuerdan algunos métodos funcionales para los cuales la

noción de distancias puede ser muy útil, por ejemplo, clasificación supervisada y no supervisada, contrastes para diferencias de medias, predicción y el concepto de función de densidad para datos funcionales.

En el Capítulo 2, se presenta una nueva semi-distancia para observaciones funcionales que generaliza la distancia de Mahalanobis para conjuntos de datos multivariantes. También se muestran las principales características de la semi-distancia de Mahalanobis funcional. Con el fin de ilustrar la aplicabilidad de esta medida de proximidad entre observaciones funcionales, se desarrollan nuevas versiones de varios procedimientos clásicos de clasificación funcional utilizando la semi-distancia de Mahalanobis funcional. Se ilustra el comportamiento de esta nueva semi-distancia con datos simulados y dos conjuntos de datos reales, lo que nos indica que los métodos de clasificación utilizados conjuntamente con la semi-distancia de Mahalanobis funcional proporcionan mejores resultados que otros procedimientos conocidos.

En el Capítulo 3 se derivan los estadísticos T^2 de Hotelling para testear la igualdad de medias en dos muestras independientes procedentes de dos distribuciones funcionales. Los estadísticos que se proponen están basados en la semi-distancia de Mahalanobis funcional y, bajo determinadas condiciones, sus distribuciones asintóticas son chi-cuadrado, sin tener en cuenta la distribución de partida de las muestras aleatorias funcionales. Se proporciona el vínculo entre los estadísticos T^2 obtenidos y estadísticos basados en la semi-distancia de componentes principales funcionales. El comportamiento de todos estos estadísticos se analiza mediante un extenso estudio de Monte Carlo y el análisis de un conjunto de datos reales recogidos en climatología. Los resultados parecen indicar que los estadísticos T^2 de Hotelling para la comparación de dos muestras superan en términos de potencia a aquellos basados en la semi-distancia de componentes principales funcionales.

Finalmente, el Capítulo 4 contiene un resumen y algunas posibles líneas de investigación futuras del trabajo presentado en esta tesis.

CHAPTER 1

Introduction and background

At the present time, there are a number of situations in diferent fields of applied sciences such as chemometrics, economics, image analysis, medicine, meteorology and speech recognition, among others, where it can be assumed that the observed data are points belonging to functions defined over a given set, $T = [a, b] \subset \mathbb{R}$. Functional data analysis (FDA) deals with such kind of observations. In practice, the values of the functions are available only at a finite number of points and, as a general rule, functional samples may contain less functions than evaluation points. For these reasons, classical methods designed for multivariate data are no longer applicable. Therefore, it is not convenient to treat functional data as multivariate data, and, consequently, there is a need to develop special techniques for this type of data. There are several methodologies for FDA being the most popular the one based on the use of basis functions such as Fourier and splines, see Ramsay and Silverman [48]. Alternatively, other procedures, such as the nonparametric approach proposed by Ferraty and Vieu [20], do not require the knowledge of the explicit form of the functions. Horváth and Kokoszka [29] review some

recent developments on inference for functional data, whereas Cuevas [11] presented a partial overview of the state of art in FDA theory. In a conceptual sense, functional data are intrinsically infinite dimensional and also the measurements within one curve display high correlation.

This dissertation is focus in $L^2(T)$, the infinite dimensional space of functions η defined on T satisfying $\int_T \eta^2(s)\,ds < \infty$. The space $L^2(T)$ is a separable Hilbert space with inner product $\langle \eta, \kappa \rangle = \int_T \eta(s)\,\kappa(s)\,ds$, for $\eta, \kappa \in L^2(T)$. Let χ be a functional random variable defined on $L^2(T)$, $\chi(t)$ being the function evaluated at point $t \in T$. We assume that χ has a functional mean, denoted by μ_χ, such that $\mu_\chi(t) = E[\chi(t)]$, for $t \in T$, and a covariance function, denoted by c_χ, such that $c_\chi(t,s) = E[(\chi(t) - \mu_\chi(t))(\chi(s) - \mu_\chi(s))]$, for $t, s \in T$. The covariance function c_χ makes possible to introduce the covariance operator of χ, denoted by Γ_χ, that transforms any $\eta \in L^2(T)$ into a new function defined on $L^2(T)$ given by:

$$\Gamma_\chi(\eta)(t) = \int_T c_\chi(t,s)\,\eta(s)ds, \qquad \text{for all} \quad t \in T. \tag{1.0.1}$$

Note that Γ_χ plays the same role in the functional framework as the covariance matrix in the multivariate context. For convenience in future developments, we express the covariance operator in terms of the inner product in $L^2(T)$ as,

$$\Gamma_\chi(\eta) = E[\langle \chi - \mu_\chi, \eta \rangle (\chi - \mu_\chi)].$$

In Figure 1.1, we present some examples of functional data considered in this dissertation:

1. First, we consider 77 near-infrared absorbance spectra of meat samples, with high fat content, measured at a common discretized set of 100-channel absorbance spectrum in the wavelength range 850-1050 nm. Those curves represent a part of the Tecator dataset that is available at http://lib.stat.cmu.edu/datasets/tecator.

2. As another example of functional data, we present 100 log-periodograms corresponding to the phoneme "ao" in the first vowel of "water". Those 100 log-periodograms belong to the Phoneme dataset described in Ferraty and Vieu [20].

3. The third case is the daily temperature records of 15 Eastern weather stations of Canada over 365 days. Those curves have been profusely analyzed in the literature of FDA, see Ramsay and Silverman [48] and Zhang and Chen [63], for instance.

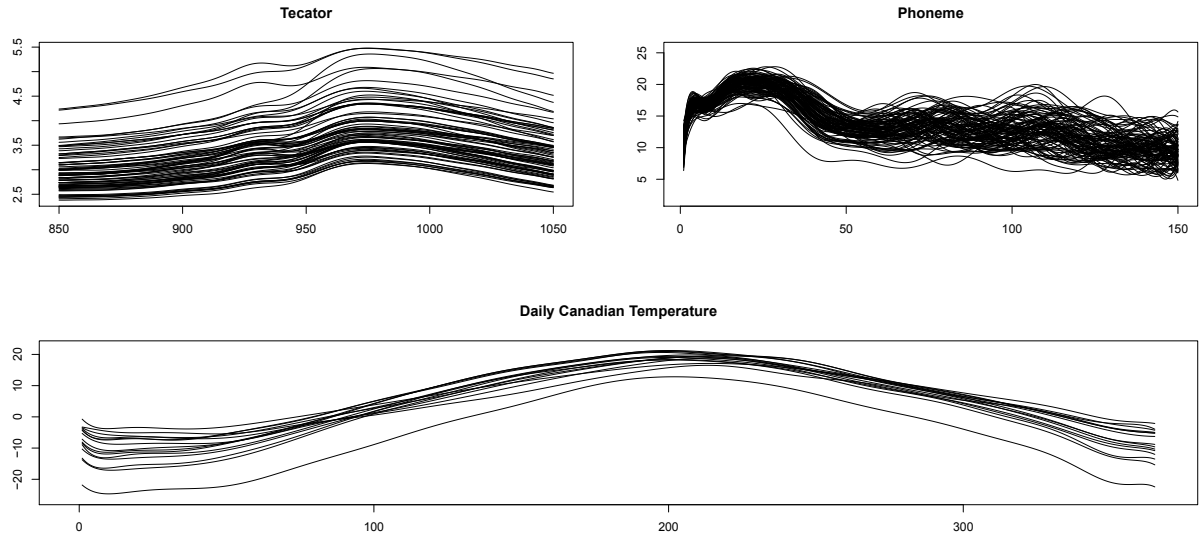


Figure 1.1: Three real functional datasets: near-infrared absorbance spectra of meat samples having high fat content (top left), log-periodograms of the phoneme "ao"(top right) and daily temperature records of Eastern weather stations of Canada (bottom)

Once the type of data that we will use in this dissertation has been illustrated, we introduce the main theoretical aspects in FDA which are necessary to understand the contributions of the following chapters. Then, in Section 1.1, we review how to build functional data from discrete observations. In Section 1.2, the notions of functional principal components are defined. In Section 1.3, we present a brief historical summary of distances in the multivariate context and recall some distances and semi-distances proposed in the literature of FDA. Finally, in Section 1.4 we present some functional methods based on distance.

1.1 FDA and basis representation

In practice, functions are usually observed with noise and are not observed continuously over all the points of T. Usually, a functional dataset has the form $\{\chi_i^*(t_{i,j}) : i = 1, \ldots, n \text{ and } j = 1, \ldots, J_i\}$, for $t_{i,j} \in T$, where n is the number of observed curves and J_i is the number of observations of the noisy curve χ_i^* at points $t_{i,1}, \ldots, t_{i,J_i}$. The number of observation points and their locations may vary over observed curves. Thus, the first step in FDA is to reconstruct the functional data from their discrete observations. One option to obtain closed form expressions of the set of functional data is to use basis functions, which is the approach taken in this thesis. This procedure consists of obtaining the coordinates of the projection of the functions in some functional sub-space of finite dimension. In general, a basis is a system of functions, denoted by ϕ_m, $m = 1, 2, \ldots$, orthogonal or not, such that, $\chi_i^*(t)$, for $i = 1, \ldots, n$, can be fairly well approximated with:

$$\chi_i(t) = \sum_{m=1}^{M} \beta_{im}\phi_m(t), \tag{1.1.1}$$

where β_{im}, for $m = 1, \ldots, M$, are the coefficients of the expansion. One of the advantages of this approach is that instead of storing all the data points, one stores the coefficients of the expansion, i.e., the β_{im}. We generally choose M so that the plotted functional objects resemble original data but with smoothing which eliminates the most obvious noise. To implement this methodology, the choice of the basis is also important and must be done according to the characteristics of the data. The function basis $\{\phi_m\}_m$ more common in applications are the classical Fourier basis and B-spline basis. For periodic or nearly periodic datasets, Fourier basis is an adequate choice. For nonperiodic datasets, B-spline basis are typically used. More basis functions for applications are presented in Ramsay and Silverman [48]. We only present in this chapter the two basis used in the thesis.

- Fourier series are useful for extremely stable functions which means functions with no strong local features and a roughly constant curvature. They are inappropriate

for functions with discontinuities or low order derivatives. The orthonormal version of the Fourier basis is given by

$$\phi_0 = \frac{1}{\sqrt{P}},\ \phi_{2r-1}(t) = \sqrt{\frac{P}{2}} sin(rwt),\ \phi_{2r}(t) = \sqrt{\frac{P}{2}} cos(rwt),\ r = 1, 2, \ldots,$$

where $P = \frac{2\pi}{w}$ is the period, i.e., the length of the interval T. An important feature of this type of basis is its easy differentiability. To define a Fourier basis system, the number of basis functions M and the period P are required.

- B-spline basis is a basis of piecewise polynomial functions defined in a recursive way. Spline coefficients are fast to compute and B-splines form a very flexible system that provide a good approximation with a relatively small M. For built a B-splines basis, the interval is divided into L subintervals separated by values ξ_l, $l = 1, \ldots, L-1$, called breakpoints or knots. Over any subinterval, the spline function is a polynomial of fixed degree or order. The term degree is used to refer the highest power in the polynomial. The order of a polynomial is the number of constants required to define it, and is one more than its degree. The number of parameters required to define a spline function is the order of the polynomial segments plus the number of interior knots.

Fourier and B-splines basis are different ways of representing functional data depending on the kind of observations we are working with. When estimating the coefficients of the basis representation, we use a smooth approximation method as least squares after choosing an appropriate basis; that is, the coefficients of the expansion are estimated by minimizing:

$$\left(\sum_{j=1}^{J_i} \left[\chi_i^* \left(t_{i,j} \right) - \sum_{m=1}^{M} \beta_{im} \phi_m \left(t_{i,j} \right) \right]^2 \right)^{1/2}.$$

Once the observed dataset $\{\chi_i^* (t_{i,j}) : i = 1, \ldots, n \text{ and } j = 1, \ldots, J_i\}$ is smoothed, we work with the smoothed functional sample $\{\chi_i (t) : i = 1, \ldots, n\}$ given in (1.1.1). More information about this topic can be found in Section 3.4 of Ramsay and Silverman [48].

1.2 Functional principal components

Principal components is a fundamental notion very useful for analyzing functional data. We present in this section the basic ideas of principal component analysis and its characterization in terms of the eigenvalues and eigenfunctions of the covariance operator of a functional variable. To motivate the concepts in the functional setting, we begin with the corresponding to the multivariate case, and then move to the infinite dimensional space.

The Principal Component Analysis (PCA) was introduced in 1901 by Pearson and developed independently in 1933 by Hotelling. It is a technique related to exploratory data analysis which aims to reduce the number of variables of a multivariate dataset. More specifically, the PCA provides an adequate representation of the information of the data in a small number of variables obtained as linear combinations of the originals. Let $\mathbf{x} = (x_1, \ldots, x_p)'$ be a continuous random variable defined in $\mathbb{R}^p$ with mean vector $\mathbf{m_x} = (m_{x_1}, \ldots, m_{x_p})'$ and positive definite covariance matrix $\mathbf{C_x}$. A linear combination of the centered components of $\mathbf{x}$ is given by:

$$s = \mathbf{v}'(\mathbf{x} - \mathbf{m_x}) = \sum_{j=1}^{p} v_j(x_j - m_{x_j}),$$

where $\mathbf{v} = (v_1, \ldots, v_p)'$ is the vector of weight coefficients applied to the components of the variable $\mathbf{x}$. In principal components, the weights are chosen in order to show types of variations that are strongly represented in the variables. More specifically, the principal components are linear combinations of the components of the variable $\mathbf{x}$ that maximize the variance of the projected variable subject to the Euclidean norm of the weights is 1. The procedure for obtaining the principal components is the following:

1. Find the weight vector $\mathbf{v}_1 = (v_{11}, \ldots, v_{p1})'$ such that the linear combination,

   $$s_1 = \sum_{j=1}^{p} v_{j1}(x_j - m_{x_j}) = \mathbf{v}_1'(\mathbf{x} - \mathbf{m_x}),$$

   has maximum variance $Var(s_1)$ subject to $\|\mathbf{v}_1\|^2 = 1$. The solution provided that $\mathbf{v}_1$ is the eigenvector of the matrix $\mathbf{C_x}$ associated with the largest eigenvalue, a_1. Furthermore, $Var(s_1) = a_1$.

2. Then, a new weight vector $\mathbf{v}_2 = (v_{12}, \ldots, v_{p2})'$ is calculated for a new linear combination $s_2 = \mathbf{v}_2'(\mathbf{x} - \mathbf{m_x})$, where $Var(s_2)$ is maximum subject to $\|\mathbf{v}_2\|^2 = 1$ and such that $\mathbf{v}_1'\mathbf{v}_2 = 0$. Then, $\mathbf{v}_2$ is the eigenvector of the matrix $\mathbf{C_x}$ related to the second largest eigenvalue, a_2. Furthermore, $Var(s_2) = a_2$.

3. The third and subsequent steps consist of repeating the previous step until obtaining the p principal components, that is the limit of the number of variables. The other weight vectors are the eigenvectors associated to the eigenvalues of $\mathbf{C_x}$ ordered according to the magnitude of the corresponding eigenvalues. The variance of these components corresponds to the eigenvalues of $\mathbf{C_x}$.

The new variables s_m are often called scores of the principal components. Let $\mathbf{v}_1, \ldots, \mathbf{v}_p$ be the eigenvectors of the covariance matrix $\mathbf{C_x}$ associated with positive eigenvalues $a_1 \geq \cdots \geq a_p > 0$, and let $\mathbf{V}$ be the $p \times p$ matrix whose columns are the eigenvectors of $\mathbf{C_x}$, i.e., $\mathbf{V} = [\mathbf{v}_1| \cdots |\mathbf{v}_p]$. The vector of principal component scores given by $\mathbf{s} = \mathbf{V}'(\mathbf{x} - \mathbf{m_x})$, is a multivariate random variable with zero mean vector and diagonal covariance matrix $\mathbf{A} = diag(a_1, \ldots, a_p)$. As a consequence, $\mathbf{x}$ can be written in terms of the principal component scores in the following way:

$$\mathbf{x} = \mathbf{m_x} + \mathbf{V}\mathbf{s}. \tag{1.2.1}$$

Additionally, the singular value decomposition of $\mathbf{C_x}$, i.e., $\mathbf{C_x} = \mathbf{V}\mathbf{A}\mathbf{V}'$, that allows the inverse of $\mathbf{C_x}$ to be written in terms of $\mathbf{V}$ and $\mathbf{A}$ as

$$\mathbf{C_x}^{-1} = \mathbf{V}\mathbf{A}^{-1}\mathbf{V}', \tag{1.2.2}$$

and as we will see in the next section, the inverse of the covariance matrix is useful to define the Mahalanobis distance.

Once we have reviewed the concept of principal components for multivariate data, we wil see how the components are obtained in the functional context.

Let χ be a functional random variable defined on $L^2(T)$ with mean function μ_χ and covariance operator Γ_χ as defined in (1.0.1). If $E\left[\|\chi\|_2^2\right]$ is finite, where $\|.\|_2 =$

$\langle .,. \rangle^{1/2}$ denotes the usual norm in $L^2(T)$, then Γ_χ is a compact operator, see Mas [42]. Consequently, there exists a sequence of non-negative eigenvalues of Γ_χ, denoted by $\lambda_1 > \lambda_2 > \cdots$, where $\sum_{k=1}^{\infty} \lambda_k < \infty$, and a set of orthonormal eigenfunctions of Γ_χ, denoted by $\psi_1, \psi_2, \ldots$ such that $\Gamma_\chi(\psi_k) = \lambda_k \psi_k$, for $k = 1, 2, \ldots$

Functional principal components analysis works similarly to the multivariate case. In the first step, it seeks a weight function φ_1 that maximizes the variance of projection $\theta_1 = \langle \varphi_1, \chi - \mu_\chi \rangle$ subject to $\|\varphi_1\|^2 = \int \varphi_1(t)^2 dt = 1$. The solution is that the weight function is the eigenfunction of the covariance operator Γ_χ associated to the eigenvalue of greater magnitude, i.e, ψ_1. Moreover, the variance of the linear combination, called scores as in the multivariate case, is the value of the eigenvalue, i.e, $Var(\theta_1) = \lambda_1$. Next, the algorithm follows the same steps as in the multivariate case. Thus, the solutions of the above algorithm are the eigenfunctions of Γ_χ. The set of eigenfunctions $\psi_1, \psi_2, \ldots$ form an orthonormal basis in $L^2(T)$ that allows Γ_χ to be written as

$$\Gamma_\chi(\eta) = \sum_{k=1}^{\infty} \lambda_k \langle \psi_k, \eta \rangle \psi_k. \tag{1.2.3}$$

The well-known Karhunen-Loève expansion of χ (see Hall and Housseini-Nassab, [24]) can be written in terms of the elements of the basis as

$$\chi = \mu_\chi + \sum_{k=1}^{\infty} \theta_k \psi_k, \tag{1.2.4}$$

where $\theta_k = \langle \chi - \mu_\chi, \psi_k \rangle$, for $k = 1, 2, \ldots$ are the functional principal component scores of χ. Note that θ_k, for $k = 1, 2, \ldots$ are uncorrelated random variables with zero mean and variance λ_k since $\psi_1, \psi_2, \ldots$ are orthonormal.

In practice, we have to estimate the functional principal component scores from the sample curves. Therefore, as mentioned in Section 1.1, the observed dataset is smoothed and then we work with the smoothed functional sample. First, we estimate the functional mean μ_χ with the sample functional mean,

$$\widehat{\mu}_\chi = \frac{1}{n} \sum_{i=1}^{n} \chi_i, \tag{1.2.5}$$

and the covariance operator Γ_χ with the sample covariance operator such that, for any $\eta \in L^2(T)$:

$$\widehat{\Gamma}_\chi(\eta) = \frac{1}{n-1}\sum_{i=1}^{n} \langle \chi_i - \widehat{\mu}_\chi, \eta \rangle (\chi_i - \widehat{\mu}_\chi). \tag{1.2.6}$$

Eigenfunctions and eigenvalues of the covariance operator Γ_χ can be approximated with those of $\widehat{\Gamma}_\chi$, leading to estimates $\widehat{\psi}_1, \widehat{\psi}_2, \ldots$ and $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots$, respectively. Additionally, the functional principal component scores corresponding to χ_i, i.e., $\theta_{i,k} = \langle \chi_i - \mu_\chi, \psi_k \rangle$, are estimated with

$$\widehat{\theta}_{i,k} = \left\langle \chi_i - \widehat{\mu}_\chi, \widehat{\psi}_k \right\rangle,\ k = 1, 2, \ldots \tag{1.2.7}$$

Computation of eigenvalues and eigenfunctions of covariance operators and functional principal component scores are described in Section 8.4 of Ramsay and Silverman [48]. Specifically, most of the inner products (integrals) that are involved in these computations can be computed exactly.

To illustrate the functional principal components concept, we use data of the mean daily temperature at 35 different locations in Canada averaged over 1960 to 1994. Figure 1.2 shows the four eigenfunctions associated to the four eigenvalues of greatest magnitude. As noted in Ramsay and Silverman [48], the first eigenfunction is positive throughout the whole year but the weight of the winter temperatures is about four times higher than that of the summer temperatures. This means that the greatest variability between weather stations will be found by heavily weighting winter temperatures, with only a light contribution from the summer days. In summary, we can say that the climate of Canada is most variable in the wintertime. Moreover, the percentage 88.8% indicates that this type of variation strongly dominates all other types of variation.

The second eigenfunction is not as important as the first because ψ_2 is orthogonal to ψ_1 and only explains a 8.4% of the total variation. This eigenfunction consists of a positive contribution for the winter period and a negative contribution for the summer period, therefore corresponding to a measure of uniformity of temperature through the year. The third and fourth eigenfunctions are orthogonal to the first two and the varia-

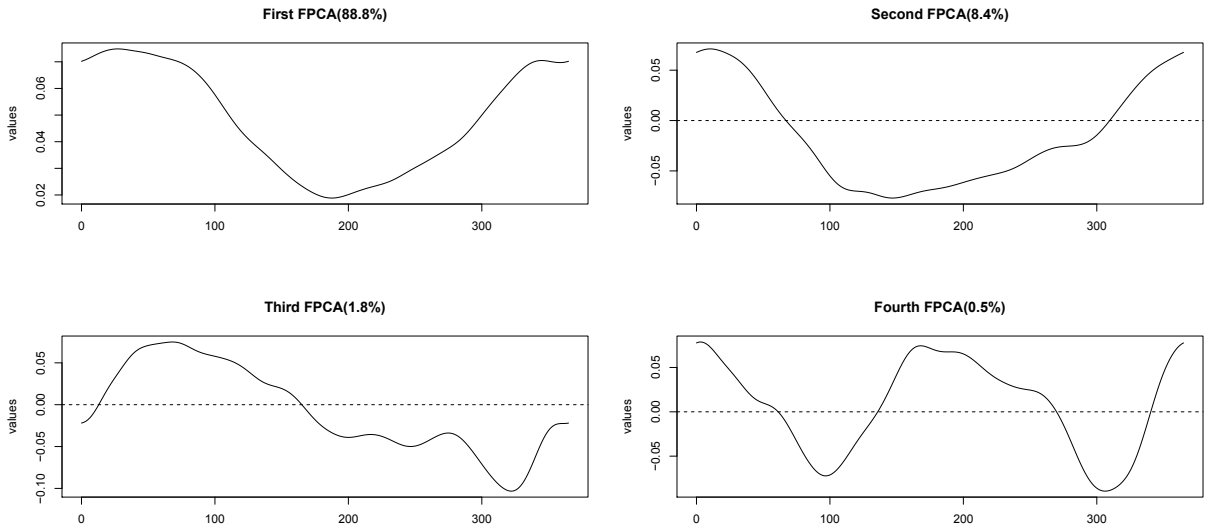


Figure 1.2: First four principal component curves of the Canadian temperature data. The percentages indicate the amount of total variation accounted for by each principal component.

tion proportions represent small proportions. These last functions are more difficult to interpret.

1.3 Distances for Functional Data

It is well known that usual multivariate methods are not usually well suited for functional datasets; however many multivariate techniques have inspired advances in FDA. The introduction of the notion of distance for functional data represents an example. In order to understand the main contribution of the thesis, that is, the definition of a new distance for functional data, this section begins summarizing the concept of distance in the multivariate context.

In the mathematical field, a distance is a function (an association rule) such that each pair of objects is associated with a nonnegative real number that satisfies certain conditions. The family of distances mostly used in $\mathbb{R}^p$ is the family of Minkowski's distances (in particular the L_1, L_2 and L_∞ distances). Naturally, these distances have

been extended to the multivariate context. Let $\mathbf{x} = (x_1, \ldots, x_p)'$ be a continuous random variable defined in $\mathbb{R}^p$ with mean vector $\mathbf{m_x} = (m_{x_1}, \ldots, m_{x_p})'$ and positive definite covariance matrix $\mathbf{C_x}$. The Minkowski's distance between $\mathbf{x}$ and $\mathbf{m_x}$ is defined by:

$$d_r(\mathbf{x}, \mathbf{m_x}) = \left(\sum_{j=1}^{p} \left| x_j - m_{x_j} \right|^r \right)^{\frac{1}{r}},$$

where r is a positive number. In particular, the case in which $r = 2$ leads to the Euclidean distance, which is one of the most important distances in Statistics. In this particular case, the above definition reduces to the following expression:

$$d_2\left(\mathbf{x}, \mathbf{m_x}\right) = \left[\left(\mathbf{x} - \mathbf{m_x}\right)' \left(\mathbf{x} - \mathbf{m_x}\right) \right]^{1/2}.$$

The Euclidean distance is invariant under orthogonal transformations. However it is strongly affected by changes of scale of the variables. To reduce this impact, an alternative distance can be considered, known as the weighted Euclidean distance, given by:

$$d_{\boldsymbol{W}}\left(\mathbf{x}, \mathbf{m_x}\right) = \left[\left(\mathbf{x} - \mathbf{m_x}\right)' \boldsymbol{W} \left(\mathbf{x} - \mathbf{m_x}\right) \right]^{1/2},$$

where $\boldsymbol{W}$ is a diagonal matrix whose elements are nonnegative real numbers $w_1, \ldots, w_p$. These weights can be defined in several ways but the most common is to take $w_j = 1/\sigma_j^2$, where $\sigma_1^2, \ldots, \sigma_p^2$ are the elements of the main diagonal of $\mathbf{C_x}$, which lead to the Pearson distance given by:

$$d_{\boldsymbol{P}}\left(\mathbf{x}, \mathbf{m_x}\right) = \left[\left(\mathbf{x} - \mathbf{m_x}\right)' \mathbf{C}_0^{-1} \left(\mathbf{x} - \mathbf{m_x}\right) \right]^{1/2},$$

where $\mathbf{C}_0 = diag(\sigma_1^2, \ldots, \sigma_p^2)$ represents the diagonal matrix containing the variances of $\mathbf{x}$. Therefore, the Pearson distance is a particular case of weighted Euclidean distance that is invariant to changes of scale.

As it can be seen, none of the previous multivariate distances take into account the correlations between the components of the variable $\mathbf{x}$. In order to include the correlations, Prasanta Chandra Mahalanobis introduced the Mahalanobis distance (MD) in order to compare the morphological measures of races in India, see Mahalanobis [39].

The MD is the most important distance in statistics and takes into account the covariance among the variables in such way that the problems of scale and correlation inherent in the Euclidean distance disappear.

Specifically, the MD between $\mathbf{x}$ and $\mathbf{m_x}$ is given by:

$$\begin{aligned} d_M\left(\mathbf{x},\mathbf{m_x}\right) &= \left\langle \mathbf{C_x}^{-1/2}\left(\mathbf{x}-\mathbf{m_x}\right), \mathbf{C_x}^{-1/2}\left(\mathbf{x}-\mathbf{m_x}\right)\right\rangle_E^{1/2} = \\ &= \left[\left(\mathbf{x}-\mathbf{m_x}\right)'\mathbf{C_x}^{-1}\left(\mathbf{x}-\mathbf{m_x}\right)\right]^{1/2}. \end{aligned} \tag{1.3.1}$$

We can also define the MD between more statistical objects. For example, the MD between two random variables $\mathbf{x}_1$ and $\mathbf{x}_2$ with the same nonsingular covariance matrix $\mathbf{C_x}$ is defined as:

$$d_{\boldsymbol{M}}\left(\mathbf{x}_1,\mathbf{x}_2\right) = \left[\left(\mathbf{x}_1-\mathbf{x}_2\right)'\mathbf{C_x}^{-1}\left(\mathbf{x}_1-\mathbf{x}_2\right)\right]^{1/2}.$$

In addition, the MD between an observation $\mathbf{x}_i$ of a sample generated by a random variable $\mathbf{x}$ and the sample mean vector based on the sample, $\widehat{\mathbf{m}}_\mathbf{x}$, is defined as:

$$d_{\boldsymbol{M}}\left(\mathbf{x}_i,\widehat{\mathbf{m}}_\mathbf{x}\right) = \left[\left(\mathbf{x}_i-\widehat{\mathbf{m}}_\mathbf{x}\right)'\widehat{\mathbf{C}}_\mathbf{x}^{-1}\left(\mathbf{x}_i-\widehat{\mathbf{m}}_\mathbf{x}\right)\right]^{1/2},\ i=1,\ldots,n,$$

where n is the number of observations of the sample generated by $\mathbf{x}$ and

$$\widehat{\mathbf{C}}_\mathbf{x} = \frac{1}{n-1}\left(\mathbf{x}_i-\widehat{\mathbf{m}}_\mathbf{x}\right)\left(\mathbf{x}_i-\widehat{\mathbf{m}}_\mathbf{x}\right)',$$

is the sample covariance matrix.

The MD in (1.3.1) is effectively a weighted Euclidean distance with the weight equal to the inverse of the covariance matrix $\mathbf{C_x}$. Therefore, as the Euclidean and Pearson distances, it shares the benefits of taking into account the measurement scale of the variables but also consider the correlation among them. Therefore, the MD is invariant under nonsingular linear transformations of the variables; in particular, changes of scale. In this sense, in several scenarios, d_M is more appropriate than d_2 and d_P. Finally, it is well known that if the multivariate random variable $\mathbf{x}$ has a multivariate Gaussian distribution, then it is easy to see that $d_M^2(\mathbf{x},\mathbf{m_x})$ has a χ_p^2 distribution and, consequently, $E\left[d_M^2(\mathbf{x},\mathbf{m_x})\right] = p$ and $V\left[d_M^2(\mathbf{x},\mathbf{m_x})\right] = 2p$.

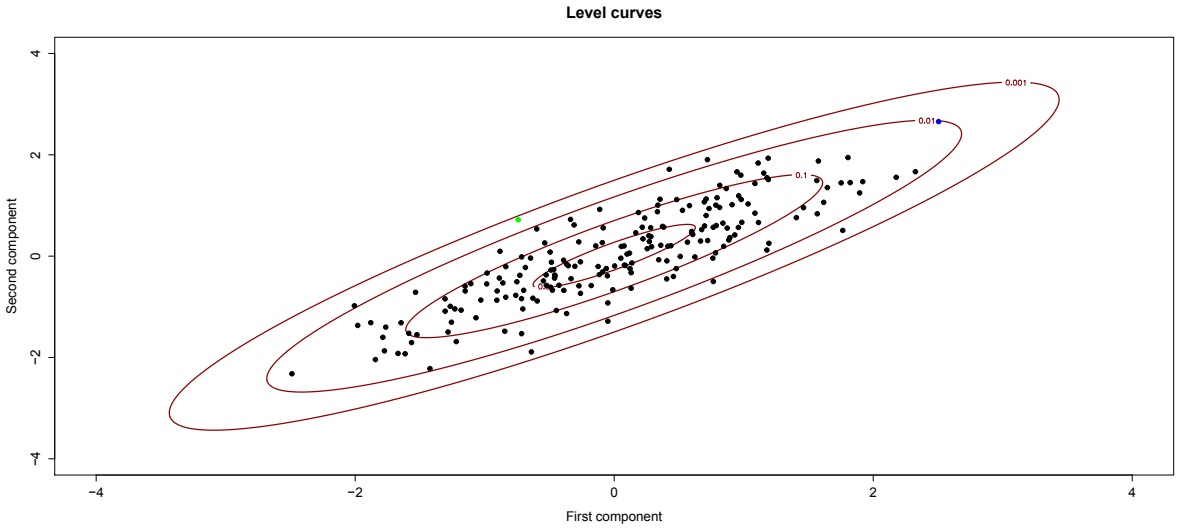


Figure 1.3: Dataset generated.

To compare the Mahalanobis distance with the Euclidean and Pearson distances using correlated data, we generate 200 random observations from a bivariate normal distribution with mean $\mathbf{m_x} = \binom{0}{0}$ and covariance matrix:

$$\mathbf{C_x} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$

and we obtain the level curves as shown in Figure 1.3. The points in each level curve form an ellipsoid. On the other hand, Figure 1.4 shows the values of the Euclidean, Pearson and Mahalanobis distances for the 200 observations. In this figure, we can see the observation that reaches greater Mahalanobis distance (drawn in green) is different from the observation with the greatest Euclidean and Pearson distances, which coincide in this case (shown in violet and blue). In other words, we can say that the green point is further from the center of the distribution than the blue using the Mahalanobis distance. The contradiction related to the farthest point measured with the different distances can be justified from the statistical standpoint. With the Mahalanobis distance the shape of the data distribution (an ellipsoid) plays a role, so the green point is actually located farther because it is located on the level curve 0.001 whereas the blue point is located on the level curve 0.01.

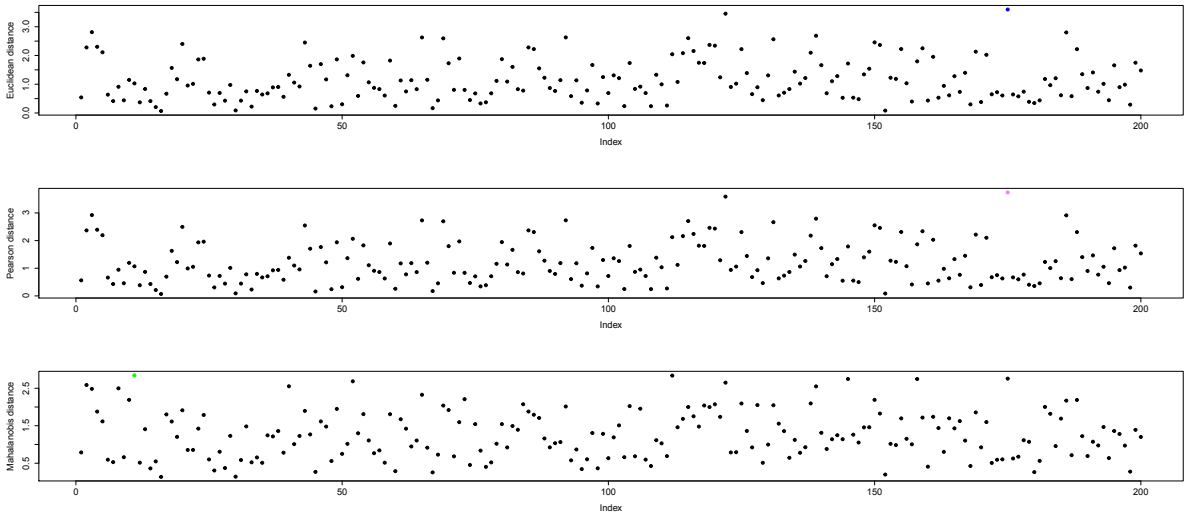


Figure 1.4: Euclidean, Pearson and Mahalanobis distances for the simulated data.

For later developments, it is important to note that the Mahalanobis distance can be rewritten in terms of the principal component scores of $\mathbf{x}$. Using (1.2.1) and (1.2.2), the Mahalanobis distance in (1.3.1) is written as

$$d_M\left(\mathbf{x}, \mathbf{m}_{\mathbf{x}}\right) = \left(\mathbf{s}'\mathbf{A}^{-1}\mathbf{s}\right)^{1/2} = \left(\mathbf{z}'\mathbf{z}\right)^{1/2}, \tag{1.3.2}$$

where $\mathbf{z} = \mathbf{A}^{-1/2}\mathbf{s}$ is the random vector of the standardized principal component scores. In other words, the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{m}_{\mathbf{x}}$ can be written as the Euclidean norm of the standardized principal component scores.

As mentioned previously, the notion of multivariate distance has been extended to functional data. It is usually assumed that functional datasets have been generated by a functional random variable defined on a Hilbert space endowed with the distance well known as the L^2 distance. Later, Chen et al. [9] have proposed a weighted L^2 distance for functional data. Ferraty and Vieu [20] have proposed semi-metrics well adapted for sample functions, including those based on functional principal components (FPC), or partial least-squares (PLS) or those based on derivatives. Next, we summarize some distances and semi-distances for functional random variables proposed previously in the literature. As mentioned before, the Euclidean distance is often used in Statistics. This distance can be extended easily to calculate the distance between a functional random

variable χ defined in $L^2(T)$ and its functional mean μ_χ, known as L^2, as follows:

$$d_2(\chi, \mu_\chi) = \langle \chi - \mu_\chi, \chi - \mu_\chi \rangle^{1/2} = \left(\int_T (\chi(t) - \mu_\chi(t))^2 \, dt \right)^{1/2}. \tag{1.3.3}$$

Chen et al. [9] propose to use a weighted L^2 distance defined in terms of w:

$$d_w(\chi, \mu_\chi) = \left(\int_T (\chi(t) - \mu_\chi(t))^2 \, w(t) dt \right)^{1/2},$$

where $w(t)$ is the weight function that satisfies $w(t) \geq 0$ and $\int_T w(t)dt = 1$. In practice, the weight function $w(t)$ have to be fixed. In order to obtain a functional distance that could be beneficial in functional clustering analysis, functional classification analysis and group-difference tests, Chen et al. [9] propose a new weight function based on minimizing the coefficient of variation of the squared distance between functional observations and that can be obtained by means of an efficient iterative procedure.

Using the Karhunen-Loève expansion of the functional random variable χ defined in (1.2.4), the L^2 functional distance in $L^2(T)$ can be written in terms of the functional principal componen scores. Thus, (1.3.3) becomes in:

$$d_2(\chi, \mu_\chi) = \left\langle \sum_{k=1}^{\infty} \theta_k \psi_k, \sum_{k=1}^{\infty} \theta_k \psi_k \right\rangle^{1/2} = \left(\sum_{k=1}^{\infty} \theta_k^2 \right)^{1/2}. \tag{1.3.4}$$

As can be seen, the L^2 distance is the square root of an infinite sum of squared functional scores. Ferraty and Vieu [20] propose a semi-distance based on the expression (1.3.4) as follows:

$$d_2^K(\chi, \mu_\chi) = \left(\sum_{k=1}^{K} \theta_k^2 \right)^{1/2}. \tag{1.3.5}$$

The expression (1.3.5) represents the semi-distance constructed from functional principal componen scores. It provides a complete family of semi-distances in the functional space $L^2(T)$ that can be applied to functional data sets. It is also possible to construct semi-distances from the derivatives of functions. Unlike the family of semi-distances based on functional principal components, the semi-distances based on derivatives can only be used with smooth functions. They are defined as follows:

$$d_{deriv}^q(\chi, \mu_\chi) = \left(\int_T \left(\chi^{(q)}(t) - \mu_\chi^{(q)}(t) \right)^2 dt \right)^{1/2},$$

where q is the order of the derivation. If $q = 0$, the usual distance of $L^2(T)$ is obtained. This family of semi-distances are appropriate when the information to study is in the curvature of the data.

Another family of semi-distances can be considered in situations when an additional response is observed. This family is based on partial least-squares (PLS). Before going on, we recall the main idea behind the PLS method.

In the classical regression analysis, there is a response and p regressors. If some of the p regressors are highly correlated, it is well known that the estimation of the parameters of the regression model is not reliable. Several approaches have been developed to deal with this problem including ridge regression, principal components regression and PLS regression, among others. In partial least squares regression, a basis of vectors that are constructed using an iterative algorithm is obtained. In the first step, it seeks the vector that maximizes the correlation between the response variable and the p regressors. In successive iterations, vectors which fulfil the above condition are obtained in orthogonal spaces to those of previous iterations. Then, the regressors are written in terms of the new basis, and standard regression is performed with these new regressors. This procedure is somewhat similar to principal components regression (PCR). The fundamental difference between PCR and PLS is that the basis built by PLS takes into account the relationship between the regressors and the response.

As the principal component analysis, PLS regression has been extended to the functional framework, see Preda and Saporta [46]. This method can be useful for different purposes involving functional data. In particular, functional PLS allows to build a class of semi-distances essentially similar to the semi-distance based on functional principal components; that is, the semi-distance based on PLS between the functional random variable χ and its functional mean μ_χ is defined as follows:

$$d_{PLS}^{K}(\chi, \mu_\chi) = \left(\sum_{k=1}^{K} \left(\int_T (\chi(t) - \mu_\chi(t)) \gamma_k(t)\, dt \right)^2 \right)^{1/2},$$

where γ_k, for $k = 1, \ldots, K$ are the functions obtained by the functional algorithm PLS for the random function χ and the scalar predictor y.

Once we have finished reviewing the notion of distances and semi-distances for functional data, you can see that common distances frequently used in multivariate data analysis, such as the Mahalanobis distance proposed by Mahalanobis [39], have not been extended to the functional framework. To fill this gap, the main contribution of this thesis is to define the functional Mahalanobis semi-distance. Then, we use this new semi-distance to solve functional statistical problems that may require the use of distances. We focus on supervised classification and hypothesis testing and we show that the functional Mahalanobis semi-distance represents a useful tool to solve such problems.

1.4 Functional Distance-based Methods

There are many statistical problems in the multivariate analysis that require the use of distances. Those problems in the functional context would be also solved with distances. Typically, supervised and unsupervised classification, hypothesis testing, functional prediction and the definition of probability density functions for functional data can be some examples. We briefly describe these methods and we focus on the importance of the distance notion.

- Supervised functional classification is one of the problems that has been tackled using the notion of distance. In a supervised functional classification problem, there is a sample of functional observations coming from G predefined groups. The whole sample can be split into G subsamples and the aim is to classify a new functional observation in one of the G groups based on the sample information. Some methods have been developed in the literature to solve this problem. For example, Biau et al. [7] proposed to filter the training samples using a Fourier basis and then apply the k-nearest neighbor (kNN) classification to the first Fourier coefficients of the expansion. Baíllo et al. [3] derived several consistency results of the kNN procedure for a particular type of Gaussian process. Delaigle and Hall [15] considered the centroid method, which assigns the new function to the group with the closest mean. The authors propose to project the functions in a given

direction and then compute the squared Euclidean distance between the observations. Ferraty and Vieu [19] have proposed a method based on nonparametrically estimating the posterior probability that the new function is of a given class. This is typically a problem of discrimination, where it is used a semi-distance in the estimates of posterior probabilities. Alonso et al. [1] have proposed a weighted distance approach. In this thesis, we show that several simple classification procedures including kNN, the centroid method and functional Bayes classification rules can be used in conjunction with the functional Mahalanobis semi-distance as the criterion of proximity between functions to obtain reasonable classification rates. Several Monte Carlo experiments suggest that methods based on the functional Mahalanobis semi-distance lead to better classification rates than alternatives.

- The notion of functional distance is also important in hypothesis testing; in particular, in the problem of testing the equality of mean functions in two random samples independently drawn from two functional distributions. In the multivariate context, the two-sample Hotelling's T^2 statistic is frequently used to test the equality of means of two independent Gaussian random samples with the same covariance matrix which it is the multivariate analogue of the two sample t-test in the univariate case. Under the null hypothesis of equality of means, the Hotelling's T^2 statistic has a scaled F distribution. If equality of covariance matrices is not assumed, the testing issue is known as the multivariate Behrens-Fisher problem although the two-sample Hotelling's T^2 statistic is still used. In this case, several approximated scaled F distributions for the T^2 statistic under the null hypothesis have been proposed, see Rencher ([49], [50]), for instance. The common point of the two statistics, that is, assuming that the covariance matrices are equal or that they are different, is that the two-sample Hotelling's T^2 statistics are just the squared Mahalanobis distance between the sample means of both random samples.

  Few approaches have been proposed so far to test whether the mean functions of two functional samples are equal. For instance, Cuevas et al. [12] proposed an ANOVA test for comparing the means of multiple samples of functional data based on the L^2-norm. Estévez-Pérez and Vilar [17] proposed an ANOVA proce-

dure to compare the mean functions of several sets of curves that tries to avoid the power reduction due to the usual pre-processing of the data, as noted in Hall and Van Keilegom [26]. Benko et al. [6] developed bootstrap procedures for testing the equality of mean functions of two functional random samples, their functional principal components (FPCs), and their associated eigenvalues and eigenfunctions. Zhang, Peng and Zhang [62] and Zhang, Liang and Xiao [64] proposed a L^2-norm based statistic to test for the equality of mean functions of two Gaussian processes with possibly unequal covariance operators and derived the distributions of the proposed test statistic under the null hypothesis and a sequence of local alternatives. Finally, Horváth and Kokoszka [29] presented procedures for testing the equality of the means in two independent functional random samples based on the functional principal components semi-distance between the sample means of the two functional samples. The asymptotic distribution of the statistic derived in this way converges, under the null hypothesis, to weighted sums of squares of independent standard Gaussians. As alternative and trying to avoid the use of the weighted asymptotic distribution, Horváth and Kokoszka [29] also proposed a normalized version of the statistic based on the functional principal components semi-distance that has a chi-square limit. These inferential procedures were extended to the case of functional time series in Horváth et al. [30]. A common point of all these references is that they use the L^2 distance defined in Hilbert spaces in the development of their testing problems.

As mentioned previously, in the multivariate case, the two-sample Hotelling's T^2 statistic is just the squared Mahalanobis distance between the sample means of both samples. In this thesis, we derive two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance assuming either a common or a different covariance operator for the random samples following the ideas developed in the multivariate context. These statistics have asymptotically chi-squared distributions under the null hypothesis of equality of means and, contrary to the multivariate case, it is not necessary to consider the hypothesis of Gaussianity for the two populations. In particular, we show that the test statistics derived in

terms of the Mahalanobis semi-distance coincide with the normalized test statistic proposed by Horváth and Kokoszka [29], although, these authors did not consider the functional Mahalanobis semi-distance in the development of their normalized statistic. Therefore, we establish the link between the Hotelling's T^2 statistic in the multivariate and functional settings.

As for the classification problem, several Monte Carlo simulations are carried out to examine the performance of the test statistics based on the functional Mahalanobis semi-distance and the functional principal components semi-distance. The results suggest that the test statistics based on the functional Mahalanobis semi-distance clearly outperform the statistics based on the functional principal components semi-distance in terms of power, at least in the considered scenarios. The results appear to diverge from those of the simulation study found in Horváth and Kokoszka [29], who indicated that neither of the two tests statistics clearly dominates the other for their simulated Gaussian data. Additionally, the analysis of a real data example from climatology suggests that the test statistic based on the functional Mahalanobis semi-distance might be more powerful than the one based on the functional principal components semi-distance.

- The unsupervised classification or clustering problem is another area of application of functional distances and semi-distances. The main objective in unsupervised classification is to divide a set of functions (typically large) $\chi_1, \ldots, \chi_n$ in a number of classes k in such way that the elements of each class have some sort of similarity. This similarity will be defined by the algorithm and the metric used. One of the main challenges in unsupervised classification is the choice of the number of classes. Some algorithms provide suggestions for this choice. The unsupervised classification is used when there are groups in the data set being the main goal the identification of split structure among those groups in order to facilitate the understanding of the data set.

  The hierarchical and non-hierarchical approaches of the multivariate analysis have been extended to the functional case. Ferraty and View [20] have proposed a dis-

sociative hierarchical method using a non-parametric approach. The methodology involves iterative partition in increasingly heterogeneous groups, where this heterogeneity is measured in terms of the proximity to several functional centrality measures (mean, median or mode). The functional distances and semi-distances play a fundamental role not only in the calculation of the functional median and mode, but also in the measure of closeness between the observed functions and these centrality measures. On the other hand, Chiou and Li [10] have proposed a functional version of the k-means algorithm, called k-centers FC. This algorithm is based on an initial partition of the functions in a certain number of groups using the cluster analysis of the functional principal components scores of each observed function. Then, the curves are reclassified depending on the L_2 distance between the observed functions and a function that is considered as the center of each group. Therefore, the notion of distance is again important in this statistical problem.

- Another statistical topic where the distances and semi-distances are crucial is functional prediction. There are many situations in which one may wish to study the relation between two variables, with the main purpose to be able to predict new values of one of them given the other one. This prediction problem can occur when some of the variables are functional. Let $(\chi_i, Y_i)_{i=1,\dots,n}$ be n independent pairs, identically distributed as $(\boldsymbol{\chi}, Y)$, where $\boldsymbol{\chi}$ is a functional random variable defined in the functional space E, and Y is a scalar random variable defined in $\mathbb{R}$. It is assumed that E is endowed with a distance o semi-distance, d. Then, given a function χ generated by $\boldsymbol{\chi}$, we would like to predict the value of the scalar response Y using the regression nonlinear operator r defined by $r(\chi) = E\left(Y|\boldsymbol{\chi} = \chi\right)$. Therefore, if $\hat{r}$ is an estimator of r, a prediction of the value Y is obtained from:

  $$\hat{y} = \hat{r}(\chi).$$

  The estimate of r can be obtained using the functional kernel regression estimator:

  $$\hat{r}(\chi) = \frac{\sum_{i=1}^{n} Y_i K\left(h^{-1} d(\chi, \chi_i)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(\chi, \chi_i)\right)},$$

  where K is an asymmetrical kernel and h (depending on n) is a strictly positive real. As can be seen in Härdle [27], the estimator of r is the functional extension of

the Nadaraya-Watson estimate for finite dimensional nonparametric regression, see Nadaraya [44] and Watson [59]. The main novelty comes from the use of a semi-distance d which measures the proximity between the function χ and the observed curves.

- Finally, the concept of density function of a random function is another notion that has been developed using the notion of distance. In general, the probability density function does not exist for functional data. Delaigle and Hall [14] show that it is possible to develop the notion of density when functional data are considered in the space determined by the eigenfunctions of principal component analysis. This leads to a surrogate for density function of functional variables defined as follows:

$$P\left(d_2\left(\boldsymbol{\chi}, \chi\right) \leq h\right), \tag{1.4.1}$$

where $d_2\left(\boldsymbol{\chi}, \chi\right)$ denotes the L_2 distance between the functional variable $\boldsymbol{\chi}$ and any function χ. Delaigle and Hall [14] show that the probability in (1.4.1) can be written in terms of the average value of the logarithms of the densities of the distributions of principal components for a finite dimension. This alternative density is estimable easily derived from a data set. The possible implications of this result are of great interest in a large number of problems including the outlier detection, the definition of robust estimators of location and scale, the classification and cluster analysis of functional data, etc...

1.5 Structure of the Thesis

Once we have reviewed the main notions in FDA necessary to follow the main contributions of the thesis, we describe the structure of this document which is divided into four chapters.

In the current chapter we have reviewed some FDA issues, presented three real functional datasets and outlined the notion of functional principal components. Additionally, we have presented a brief historical summary of distances in the multivariate context and

how the concept of distance has been extended to FDA. Finally, we have recalled some functional methods for which the notion of distance can be very useful, e.g., supervised and unsupervised classification, hypothesis testing, prediction and the concept of density function for functional data.

The contributions of this dissertation are developed in Chapters 2 and 3. In Chapter 2, we present a new semi-distance for functional observations that generalizes the Mahalanobis distance for multivariate datasets to the functional framework. The main characteristics of the functional Mahalanobis semi-distance are shown. In order to illustrate the applicability of this measure of proximity between functional observations, new versions of several well known functional classification procedures are developed using the functional Mahalanobis semi-distance. A Monte Carlo study and the analysis of two real data examples indicate that the classification methods used in conjunction with the functional Mahalanobis semi-distance give better results than other well-known functional classification procedures. The results presented in Chapter 2 are published in the paper by Galeano et al. [23] available online at http://amstat.tandfonline.com/doi/abs/10.1080/00401706.2014.902774.

In Chapter 3, we derive two-sample Hotelling's T^2 statistics for testing the equality of means in two samples independently drawn from two functional distributions. The statistics that we propose are based on the functional Mahalanobis semi-distance and, under certain conditions, their asymptotic distributions are chi-squared, regardless the distribution of the functional random samples. We provide the link between the two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance and statistics based on the functional principal components semi-distance. The behavior of all these statistics is analyzed by means of an extensive Monte Carlo study and the analysis of a real data set collected in climatology. The results appear to indicate that the two-sample Hotelling's T^2 statistics outperform in terms of power those based on the functional principal components semi-distance.

Finally, some conclusions and possible future research lines are given in Chapter 4.

CHAPTER 2

The Mahalanobis distance for functional data with applications to classification

2.1 Introduction

The family of Minkowski's distances (in particular the L_1, L_2 and L_∞ distances) has been naturally extended to the functional context. However, as we commented in the Introduction, the most important statistical distance, i.e., the Mahalanobis distance (MD), has not been considered in the functional field. The first contribution of this chapter is to generalize the usual Mahalanobis distance for multivariate datasets to the functional framework. The development uses a regularized square root inverse operator in Hilbert spaces defined from the operator proposed by Mas [42] which implies that we provide a semi-distance instead of a distance but with interesting analytic properties.

As mentioned in Section 1.4, the use of distances is usual in several procedures in functional statistical problems. Specifically, most of the well-known methods for classification are distance-based. Nowadays, there is a wide variety of methods which have been developed to solve the functional classification problem. For instance, Hall et al. [25] proposed to nonparametrically estimate the probability densities of the sets

of functional principal component scores and then to estimate the posterior probability that a new function belongs to a given class using the Bayes classification rule. Leng and Müller [35] used functional logistic regression on the functional principal component scores of the training samples to classify collections of temporal gene expression curves. Biau et al. [7] proposed to filter the training samples using a Fourier basis and then apply the k-nearest neighbor (kNN) classification to the first Fourier coefficients of the expansion. Baíllo et al. [3] derived several consistency results of the kNN procedure for a particular type of Gaussian process. Delaigle and Hall [15] considered the centroid method, which assigns the new function to the group with the closest mean. James and Hastie [31] used a natural cubic spline basis plus random error to model the observations from each individual. The spline is parameterized using a basis function multiplied by a coefficient vector, which is modelled using a Gaussian distribution and the method uses the usual multivariate linear and quadratic discriminant rules on the coefficients of the cubic splines after a convenient rank-reduced analysis. Preda et al. [46] used functional PLS regression to obtain discriminant functions. Shin [56] considered an approach based on reproducing kernel Hilbert spaces. Ferraty and Vieu [19] have proposed a method based on nonparametrically estimating the posterior probability that the new function is of a given class. López-Pintado and Romo [38], Cuevas et al. [13] and Sguera et al. [55] have proposed classifiers based on the notion of data depth that are well suited for datasets containing outliers. Rossi and Villa [52] and Martín-Barragán et al. [41] have investigated the use of support vector machines (SVMs) for functional data. Wang et al. [58] have considered classification for functional data using wavelet basis functions. Epifanio [16] has developed classifiers based on shape descriptors. Finally, Alonso et al. [1] have proposed a weighted distance approach.

The second contribution of this chapter is to develop new versions of several well known functional classification procedures including kNN, the centroid method and functional Bayes classification rules using the functional Mahalanobis semi-distance. Finally, we will illustrate the performance of this new semi-distance using simulated and real data. The numerical results suggest that methods based on the functional Mahalanobis semi-distance lead to better classification rates than alternatives.

This chapter is organized as follows. Section 2.2 introduces the functional Mahalanobis semi-distance and shows some of its main characteristics. Section 2.3 reviews several classification methods for functional data and provides new approaches to these methods based on the functional Mahalanobis semi-distance. Section 2.4 analyzes the empirical properties of the procedures via several Monte Carlo experiments and illustrates the good behavior of the classification methods in conjunction with the functional Mahalanobis semi-distance through of the analysis of two real data examples. Finally, several conclusions are drawn in Section 2.5.

2.2 The functional Mahalanobis semi-distance

The goal of this section is to define the functional Mahalanobis semi-distance. In Section 2.2.1, we introduce the definitions assuming that the data are functions and then in Section 2.2.2, we give the useful tools necessary for a practical implementation of the functional Mahalanobis semi-distance when the functions are only recorded on some finite points.

2.2.1 Definitions and some characteristics

We define the functional Mahalanobis semi-distance that generalizes the usual Mahalanobis distance for multivariate datasets. Let $\mathbf{x}$ be a continuous random variable defined in $\mathbb{R}^p$ with mean vector $\mathbf{m_x}$ and positive definite covariance matrix $\mathbf{C_x}$. The MD between $\mathbf{x}$ and $\mathbf{m_x}$ is defined as the Euclidean norm of the random vector $\mathbf{C_x}^{-1/2}(\mathbf{x}-\mathbf{m_x})$ given by the relation (1.3.1). The aim is to provide a similar definition and that in order to do this the equivalent to the covariance matrix is the covariance operator. Therefore, it is necessary to define the inverse of the covariance operator, Γ_χ^{-1}. However, Γ_χ^{-1} exists only in certain circumstances. If Γ_χ^{-1} exists, it is given by:

$$\Gamma_\chi^{-1}(\zeta) = \sum_{k=1}^{\infty} \frac{1}{\lambda_k} \langle \psi_k, \zeta \rangle \psi_k,$$

where ζ is a function in the range of Γ_χ. However, Γ_χ^{-1} is an unbounded symmetric operator on $L^2(T)$ giving rise to an ill-posed problem. Since Γ_χ^{-1} is extremely irregular, Mas [42] proposed a regularized inverse operator which is a linear operator "close" to Γ_χ^{-1} and having good properties. The regularized inverse operator, denoted by Γ_K^{-1}, is defined as:

$$\Gamma_K^{-1}(\zeta) = \sum_{k=1}^{K} \frac{1}{\lambda_k} \langle \psi_k, \zeta \rangle \psi_k,$$

where K is a regularization parameter. Similarly, a regularized square root inverse operator of $\Gamma_\chi(\zeta)$ is given by:

$$\Gamma_K^{-1/2}(\zeta) = \sum_{k=1}^{K} \frac{1}{\lambda_k^{1/2}} \langle \psi_k, \zeta \rangle \psi_k, \tag{2.2.1}$$

that allows the definition of the functional Mahalanobis semi-distance between χ and μ_χ inspired by (1.3.1) as follows:

Definition 2.2.1. *Let χ be a functional random variable defined in $L^2(T)$ with mean function μ_χ and compact covariance operator Γ_χ. The functional Mahalanobis semi-distance between χ and μ_χ is defined as:*

$$d_{FM}^K(\chi, \mu_\chi) = \left\langle \Gamma_K^{-1/2}(\chi - \mu_\chi), \Gamma_K^{-1/2}(\chi - \mu_\chi) \right\rangle^{1/2}.$$

Note that we have used Γ_K^{-1} as a regularized inverse operator to define the functional Mahalanobis semi-distance. Other regularized inverse operators found in Smola and Kondor [57] may lead to alternative semi-distances for functional datasets. As previously noted in 1.3.2, the multivariate Mahalanobis distance may be expressed in terms of the principal component scores of $\mathbf{x}$. Similarly, the functional Mahalanobis semi-distance can be expressed in terms of the functional principal component scores of χ as stated in the following proposition:

Proposition 2.2.1. *The functional Mahalanobis semi-distance between χ and μ_χ can be written as*

$$d_{FM}^K(\chi, \mu_\chi) = \left(\sum_{k=1}^{K} \omega_k^2 \right)^{1/2}, \tag{2.2.2}$$

where $\omega_k = \theta_k / \lambda_k^{1/2}$, for $k = 1, \ldots, K$, are the standardized functional principal component scores of χ.

Proof. From equation 2.2.1, it is possible to write:

$$d_{FM}^{K}(\chi,\mu_\chi)=\left\langle \sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle \psi_k,\chi-\mu_\chi\right\rangle \psi_k,\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle \psi_k,\chi-\mu_\chi\right\rangle \psi_k\right\rangle^{1/2}.$$

Now, from equation 1.2.4 and ψ_k being orthonormal eigenfunctions, the previous expression leads to:

$$\begin{aligned} d_{FM}^{K}(\chi,\mu_\chi) &= \left\langle \sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left[\left\langle \psi_k,\sum_{j=1}^{\infty}\theta_j\psi_j\right\rangle \psi_k\right],\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left[\left\langle \psi_k,\sum_{j=1}^{\infty}\theta_j\psi_j\right\rangle \psi_k\right]\right\rangle^{1/2} \\ &= \left\langle \sum_{k=1}^{K}\frac{\theta_k}{\lambda_k^{1/2}}\psi_k,\sum_{k=1}^{K}\frac{\theta_k}{\lambda_k^{1/2}}\psi_k\right\rangle^{1/2} \\ &= \left(\sum_{k=1}^{K}\frac{\theta_k^2}{\lambda_k}\right)^{1/2} \\ &= \left(\sum_{k=1}^{K}\omega_k^2\right)^{1/2} \end{aligned}$$

■

Thus, the functional Mahalanobis semi-distance between χ and μ_χ is the Euclidean norm of the standardized functional principal component scores. This property provides a simple way to compute the functional Mahalanobis semi-distance in practice. Next, we extend the definition of functional Mahalanobis semi-distance to the general situation of distance between two independent and identically distributed functional random variables.

Definition 2.2.2. *Let χ_1 and χ_2 be two independent and identically distributed functional random variables defined in $L^2(T)$ with mean function μ_χ and compact covariance operator Γ_χ. The functional Mahalanobis semi-distance between the functions χ_1 and χ_2 is given by:*

$$d_{FM}^{K}(\chi_1,\chi_2)=\left\langle \Gamma_K^{-1/2}(\chi_1-\chi_2),\Gamma_K^{-1/2}(\chi_1-\chi_2)\right\rangle^{1/2}.$$

Definition 2.2.2 leads to the following proposition that will be used in Section 2.3.1:

Proposition 2.2.2. *The functional Mahalanobis semi-distance between χ_1 and χ_2 can be written as follows:*

$$d_{FM}^K(\chi_1,\chi_2)=\left(\sum_{k=1}^{K}\left(\omega_{1k}-\omega_{2k}\right)^2\right)^{1/2}, \tag{2.2.3}$$

where $\omega_{1k}=\theta_{1k}/\lambda_k^{1/2}$ and $\omega_{2k}=\theta_{2k}/\lambda_k^{1/2}$, for $k=1,2,\ldots$ are the standardized functional principal component scores of χ_1 and χ_2, respectively.

Proof. By hypothesis, the two functions χ_1 and χ_2 have the same mean function, μ_χ, and the same covariance operator, Γ_χ. Therefore, the Karhunen-Loève expansions of χ_1 and χ_2 are $\chi_1=\mu_\chi+\sum_{k=1}^{\infty}\theta_{1k}\psi_k$ and $\chi_2=\mu_\chi+\sum_{k=1}^{\infty}\theta_{2k}\psi_k$, respectively, where $\theta_{1k}=\langle\chi_1-\mu_\chi,\psi_k\rangle$ and $\theta_{2k}=\langle\chi_2-\mu_\chi,\psi_k\rangle$, for $k=1,\ldots$ are the functional principal component scores of χ_1 and χ_2, respectively. Consequently, the difference between the two functions χ_1 and χ_2 can be written as:

$$\chi_1-\chi_2=\sum_{k=1}^{\infty}(\theta_{1k}-\theta_{2k})\psi_k. \tag{2.2.4}$$

Using the equation 2.2.1, the Mahalanobis semi-distance between χ_1 and χ_2 is given by:

$$d_{FM}^K(\chi_1,\chi_2)=\left\langle\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle\psi_k,\chi_1-\chi_2\right\rangle\psi_k,\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle\psi_k,\chi_1-\chi_2\right\rangle\psi_k\right\rangle^{1/2}$$

Now, ψ_k being orthonormal eigenfunctions and by (2.2.4), the above expression can be written as:

$$\begin{aligned}
d_{FM}^K(\chi_1,\chi_2) &= \left\langle\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle\psi_k,\chi_1-\chi_2\right\rangle\psi_k,\sum_{k=1}^{K}\frac{1}{\lambda_k^{1/2}}\left\langle\psi_k,\chi_1-\chi_2\right\rangle\psi_k\right\rangle^{1/2}\\
&= \sum_{k=1}^{K}\frac{1}{\lambda_k}\left\langle\left\langle\psi_k,\sum_{j=1}^{\infty}(\theta_{1j}-\theta_{2j})\psi_j\right\rangle\psi_k,\left\langle\psi_k,\sum_{j=1}^{\infty}(\theta_{1j}-\theta_{2j})\psi_j\right\rangle\psi_k\right\rangle^{1/2}\\
&= \left(\sum_{k=1}^{K}\frac{1}{\lambda_k}\left\langle(\theta_{1k}-\theta_{2k})\psi_k,(\theta_{1k}-\theta_{2k})\psi_k\right\rangle\right)^{1/2}\\
&= \left(\sum_{k=1}^{K}(\omega_{1k}-\omega_{2k})^2\right)^{1/2},
\end{aligned}$$

where $\omega_{1k}=\theta_{1k}/\lambda_k^{1/2}$ and $\omega_{2k}=\theta_{2k}/\lambda_k^{1/2}$, for $k=1,2,\ldots$ are the standardized functional principal component scores of χ_1 and χ_2, respectively. ■

The next result shows that d_{FM}^K is indeed a functional semi-distance and therefore, it is properly defined.

Proposition 2.2.3. *Let χ_1, χ_2 and χ_3 be three independent and identically distributed functional random variables defined in $L^2(T)$ with mean function μ_χ and compact covariance operator Γ_χ. For any positive integer K, d_{FM}^K is a functional semi-distance, as it satisfies the following three properties:*

1. $d_{FM}^K(\chi_1, \chi_2) \geq 0$.
2. $d_{FM}^K(\chi_1, \chi_2) = d_{FM}^K(\chi_2, \chi_1)$.
3. $d_{FM}^K(\chi_1, \chi_2) \leq d_{FM}^K(\chi_1, \chi_3) + d_{FM}^K(\chi_3, \chi_2)$.

The proof of this proposition is straightforward in view of Proposition 2.2.2. Note that $d_{FM}^K(\chi_1, \chi_2)$ is not a functional distance because $d_{FM}^K(\chi_1, \chi_2) = 0$ if χ_1 and χ_2 have the same first K functional principal component scores, which does not imply $\chi_1 = \chi_2$. To end this subsection, the following result, shows that for a Gaussian process, the squared functional Mahalanobis semi-distance between χ and μ_χ has a chi-squared distribution.

Theorem 2.2.1. *If χ is a Gaussian process, $d_{FM}^K(\chi, \mu_\chi)^2 \sim \chi_K^2$.*

Proof. The result is trivial because, as χ is a Gaussian process, the standardized functional principal component scores, ω_k, for $k = 1, 2, \ldots$ are independent standard Gaussian random variables, see Ash and Gardner [2]. ■

2.2.2 Practical implementation

As mentioned in the Introduction of this thesis (see Section 1.1), the functions are usually observed with noise and are not observed continuously over all the points of $T = [a, b]$. Thus, calculation of the functional Mahalanobis semi-distances as defined in (2.2.2) and (2.2.3) is not possible. Recall that the observed dataset is previously smoothed using the expression (1.1.1). Then, we work with the smoothed functional sample and estimate the functional mean μ_χ and the covariance operator Γ_χ by the relations (1.2.5) and (1.2.6), respectively. Eigenfunctions and eigenvalues of the covariance operator Γ_χ can be approximated with those of $\widehat{\Gamma}_\chi$. Additionally, The functional principal component

scores corresponding to χ_i, i.e., $\theta_{i,k} = \langle \chi_i - \mu_\chi, \psi_k \rangle$, are estimated with $\widehat{\theta}_{i,k}$ as defined in (1.2.7) which allows us to define the functional Mahalanobis semi-distance between χ_i and the functional sample mean $\widehat{\mu}_\chi$ as follows:

$$d_{FM}^K(\chi_i, \widehat{\mu}_\chi) = \left(\sum_{k=1}^{K} \widehat{\omega}_{ik}^2 \right)^{1/2},$$

where $\widehat{\omega}_{ik} = \widehat{\theta}_{i,k}/\widehat{\lambda}_k^{1/2}$, for $k = 1, \ldots, K$, are the sample standardized functional principal component scores and $\widehat{\lambda}_k$ are the eigenvalues of the sample covariance operator $\widehat{\Gamma}_\chi$. Similarly, using Proposition 2.2.2, the functional Mahalanobis semi-distance between two functions of the sample, χ_i and $\chi_{i'}$, can be written as follows:

$$d_{FM}^K(\chi_i, \chi_{i'}) = \left(\sum_{k=1}^{K} (\widehat{\omega}_{ik} - \widehat{\omega}_{i'k})^2 \right)^{1/2},$$

where $\widehat{\omega}_{i'k} = \widehat{\theta}_{i',k}/\widehat{\lambda}_k^{1/2}$, for $k = 1, \ldots, K$.

The choice of the regularization parameter K is an important aspect in practice depending on the situation where the functional Mahalanobis semi-distance is used. In the case of classification, we choose the threshold value K via leave-one-out cross validation as will be seen in Section 2.4.1.

Finally, we note that we use the methods described in Ramsay and Silverman [48] and implemented in the **R** package *fda*, see Ramsay et al. [47] to carry out all the computations.

2.3 Classification with the functional Mahalanobis semi-distance

The aim of this section is to propose new procedures based on the combination of well-known functional classification methods with the functional Mahalanobis semi-distance. We consider a sample of functional observations coming from G predefined groups. Then, the whole sample can be split into G subsamples, denoted by $\chi_{g1}, \ldots, \chi_{gn_g}$, for $g = 1, \ldots, G$, respectively, where $n = n_1 + \cdots + n_G$ is the sample size of the whole dataset.

The goal is to classify a new functional observation χ_0 in one of the G groups based on the sample information.

2.3.1 The k-nearest neighbor (kNN) procedure

The k-nearest neighbor (kNN) procedure is one of the most popular methods used to perform classification in multivariate settings. Its generalization to infinite-dimensional spaces has been studied by Biau et al. [7], Cérou and Guyader [8] and Baíllo et al. [3], among others. The kNN method starts by computing the distances between the new function to classify, χ_0, and all the functions in the observed sample. Then, the method classifies χ_0 in the group which is most common among the k functional observations closest in distance to χ_0. Cérou and Guyader [8] have obtained sufficient conditions for consistency of the kNN classifier when the functional random variable takes values in a separable metric space. Additionally, Baíllo et al. [3] have shown that the optimal classification rule can be explicitly obtained for a class of Gaussian processes with triangular covariance operators.

Our proposal is to use the kNN classifier in conjunction with the functional Mahalanobis semi-distance. To this end, two different ways to compute the functional Mahalanobis semi-distance can be considered, depending on whether the covariance operator can be assumed to be the same for all the classes.

In the first way, assume that the functional means under class g, denoted by μ_{χ_g}, are different but the covariance operator, denoted by Γ_χ, is the same for all the classes. Then, the functional means, μ_{χ_g}, are estimated using the functional sample mean of the functions in class g, i.e.:

$$\widehat{\mu}_{\chi_g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \chi_{gi}, \tag{2.3.1}$$

while the common covariance operator, Γ_χ, is estimated with the within class covariance operator given by:

$$\widehat{\Gamma}_\chi(\eta) = \frac{1}{n-1} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left\langle \chi_{gi} - \widehat{\mu}_{\chi_g}, \eta \right\rangle \left(\chi_{gi} - \widehat{\mu}_{\chi_g} \right), \tag{2.3.2}$$

for any $\eta \in L^2(T)$. Let $\widehat{\psi}_1, \ldots, \widehat{\psi}_K$ and $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$ be, respectively, the eigenfunctions and eigenvalues of the sample within class covariance operator (2.3.2). Using Proposition 2.2.2, the functional Mahalanobis semi-distance between χ_0 and the functional observation χ_{gi} for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$ is given by:

$$d_{FM}^{K}(\chi_0, \chi_{gi}) = \left(\sum_{k=1}^{K} \left(\widehat{\omega}_{g0k} - \widehat{\omega}_{gik} \right)^2 \right)^{1/2}, \tag{2.3.3}$$

where $\widehat{\omega}_{g0k} = \widehat{\theta}_{g0k}/\widehat{\lambda}_k^{1/2}$ and $\widehat{\omega}_{gik} = \widehat{\theta}_{gik}/\widehat{\lambda}_k^{1/2}$, respectively, are the standardized sample functional principal component scores given by $\widehat{\theta}_{g0k} = \left\langle \chi_0 - \widehat{\mu}_{\chi_g}, \widehat{\psi}_k \right\rangle$ and $\widehat{\theta}_{gik} = \left\langle \chi_{gi} - \widehat{\mu}_{\chi_g}, \widehat{\psi}_k \right\rangle$, respectively. Under the same assumptions for the means and covariance operator of the G classes, the functional principal components (FPC) semi-distance proposed by Ferraty and Vieu [20] between χ_0 and the functional observation χ_{gi} for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$, is given by:

$$d_{FPC}^{K'}(\chi_0, \chi_{gi}) = \left(\sum_{k=1}^{K'} \left(\widehat{\theta}_{g0k} - \widehat{\theta}_{gik} \right)^2 \right)^{1/2}, \tag{2.3.4}$$

where K' is a certain threshold.

In the second way, assume that both the functional means and the covariance operators, denoted by Γ_{χ_g}, are different for classes $1, \ldots, G$. The functional means, μ_{χ_g}, are estimated using (2.3.1), while the covariance operator of each class is estimated using the functional sample covariance operator of the functions in class g, i.e.:

$$\widehat{\Gamma}_{\chi_g}(\eta) = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} \left\langle \chi_{gi} - \widehat{\mu}_{\chi_g}, \eta \right\rangle \left(\chi_{gi} - \widehat{\mu}_{\chi_g} \right), \tag{2.3.5}$$

for any $\eta \in L^2(T)$. The functional Mahalanobis semi-distance between χ_0 and the functional observation χ_{gi} for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$, is as in (2.3.3). However, here $\widehat{\omega}_{g0k} = \widehat{\theta}_{g0k}/\widehat{\lambda}_{gk}^{1/2}$ and $\widehat{\omega}_{gik} = \widehat{\theta}_{gik}/\widehat{\lambda}_{gk}^{1/2}$, respectively, where $\widehat{\theta}_{g0k} = \left\langle \chi_0 - \widehat{\mu}_{\chi_g}, \widehat{\psi}_{gk} \right\rangle$, $\widehat{\theta}_{gik} = \left\langle \chi_{gi} - \widehat{\mu}_{\chi_g}, \widehat{\psi}_{gk} \right\rangle$ and $\widehat{\psi}_{g1}, \ldots, \widehat{\psi}_{gK}$ and $\widehat{\lambda}_{g1}, \ldots, \widehat{\lambda}_{gK}$ are, respectively, the eigenfunctions and eigenvalues of the sample covariance operator (2.3.5). Also, the FPC semi-distance in this second case can be written as in (2.3.4) but considering the sample functional scores obtained with the eigenfunctions from (2.3.5).

2.3.2 The centroid procedure

The centroid procedure for functional datasets, proposed by Delaigle and Hall [15], is probably the fastest and simplest classification method for functional observations. The centroid method assigns a new function χ_0 to the class with the closest mean. Delaigle and Hall [15] considered the case of $G = 2$ classes that have different mean and a common covariance operator and proposed to project the functions in a given direction and then compute the squared Euclidean distance between the observations. More precisely, Delaigle and Hall [15] proposed to use the centroid classifier with the distance between χ_0 and the sample functional mean $\widehat{\mu}_{\chi_g}$, for $g = 1, 2$, denoted by DH, and given by:

$$d_{DH}\left(\chi_0, \widehat{\mu}_{\chi_g}\right) = \left| \sum_{k=1}^{K''} \widehat{\omega}_{0gk} \widehat{\delta}_{12k} \right|, \tag{2.3.6}$$

where K'' is a certain threshold, $\widehat{\omega}_{0gk}$ is computed as in the Section 2.3.1 assuming a common covariance operator and $\widehat{\delta}_{12k} = \left\langle \widehat{\mu}_{\chi_2} - \widehat{\mu}_{\chi_1}, \widehat{\psi}_k \right\rangle / \widehat{\lambda}_k^{1/2}$, for $k = 1, \ldots, K''$. Our proposal is to use the centroid method as in the multivariate case but with the functional Mahalanobis semi-distance. As in the kNN procedure, the method also depends on whether the covariance operators are assumed equal or not.

2.3.3 The functional linear and quadratic Bayes classification rules

In multivariate statistics, the Bayes classification rule is derived as follows. Let $\boldsymbol{x}$ be a p-dimensional continuous random variable and let $f_1, \ldots, f_G$ be the corresponding density functions of $\boldsymbol{x}$ under the G classes. Let $\pi_1, \ldots, \pi_G$ be the prior probabilities assigned to the G classes, with $\pi_1 + \cdots + \pi_G = 1$. Using Bayes' Theorem, the posterior probability that a new observation $\boldsymbol{x_0}$ generated from $\boldsymbol{x}$ comes from class g is given by:

$$P(g|\boldsymbol{x_0}) = \frac{\pi_g f_g(\boldsymbol{x_0})}{\pi_1 f_1(\boldsymbol{x_0}) + \cdots + \pi_G f_G(\boldsymbol{x_0})}, \tag{2.3.7}$$

respectively. The Bayes rule classifies $\boldsymbol{x_0}$ in the class with the largest posterior probability. In particular, if the f_g densities are assumed to be Gaussian with different means $\mathbf{m}_{\boldsymbol{x}_g}$ but identical covariance matrix $\mathbf{C}_{\boldsymbol{x}}$, this is equivalent to classifying $\boldsymbol{x_0}$ in class g if $d_M\left(\boldsymbol{x_0}, \mathbf{m}_{\boldsymbol{x}_g}\right)^2 - 2\log\pi_g$ is minimum, where $d_M\left(\boldsymbol{x_0}, \mathbf{m}_{\boldsymbol{x}_g}\right)^2 =$

$\left(\boldsymbol{x_0} - \mathbf{m}_{\boldsymbol{x}_g}\right)' \mathbf{C}_{\boldsymbol{x}}^{-1} \left(\boldsymbol{x_0} - \mathbf{m}_{\boldsymbol{x}_g}\right)$ is the squared Mahalanobis distance between $\boldsymbol{x_0}$ and $\mathbf{m}_{\boldsymbol{x}_g}$. Our proposal for functional observations is to consider a similar rule but to replace the multivariate Mahalanobis distance with the functional Mahalanobis semi-distance. Consequently, assuming different means and a common covariance operator, the new observation χ_0 is assigned to the class g if $d_{FM}^K \left(\chi_0, \widehat{\mu}_{\chi_g}\right)^2 - 2\log \pi_g$ is minimum.

Moreover, if in the multivariate case the f_g densities are assumed to be Gaussian with different means $\mathbf{m}_{\boldsymbol{x}_g}$ and different covariance matrices $\mathbf{C}_{\boldsymbol{x}_g}$, the Bayes rule classifies $\boldsymbol{x_0}$ in class g if $d_M \left(\boldsymbol{x_0}, \mathbf{m}_{\boldsymbol{x}_g}\right)^2 + \log \left|\mathbf{C}_{\boldsymbol{x}_g}\right| - 2\log \pi_g$ is minimum, where $d_M \left(\boldsymbol{x_0}, \mathbf{m}_{\boldsymbol{x}_g}\right)^2 = \left(\boldsymbol{x_0} - \mathbf{m}_{\boldsymbol{x}_g}\right)' \mathbf{C}_{\boldsymbol{x}_g}^{-1} \left(\boldsymbol{x_0} - \mathbf{m}_{\boldsymbol{x}_g}\right)$, is the squared Mahalanobis distance between $\boldsymbol{x_0}$ and $\mathbf{m}_{\boldsymbol{x}_g}$. Our proposal for functional observations is to classify the new observation χ_0 to the class G_0 if:

$$d_{FM}^K(\chi_0, \widehat{\mu}_{\chi_g})^2 + \sum_{k=1}^{K} \log(\widehat{\lambda}_{gk}) - 2\log \pi_g, \tag{2.3.8}$$

is minimum, where $\widehat{\lambda}_{gk}$, for $k = 1, \ldots, K$ are the eigenvalues of the estimated covariance operators under class g, respectively, and K is the number of eigenfunctions used to compute the functional Mahalanobis semi-distances.

It is important to note that although the proposed functional linear and quadratic classification Bayes rules have been derived using the functional Mahalanobis semi-distance, these methods are equivalent to applying the multivariate linear and quadratic Bayes rules to the first K functional principal component scores. Hall et al. [25] proposed to use the Bayes classification rule in (2.3.7) after estimating nonparametrically the density function of the functional principal component scores. However, these authors have pointed out that a computationally less expensive method is to use the multivariate quadratic Bayes classification rule which is essentially the rule given in (2.3.8). Additionally, note that the proposed functional linear and quadratic Bayes classification rules are different than those proposed by James and Hastie [31] for classifying irregularly sampled curves. The idea of the paper by James and Hastie [31] is to use a natural cubic spline basis plus random error to model the observations from each individual. Then, the spline is parameterized using a basis function multiplied by a coefficient vector. This coefficient

vector is modeled using a Gaussian distribution that allows the pooling of the observed functions to estimate the mean and covariance for each class by means of an Expectation-Maximization (EM) algorithm. In other words, the method uses the usual multivariate linear and quadratic discriminant rules on the coefficients of the cubic splines of each curve after a convenient rank-reduced analysis and therefore, the method proposed by James and Hastie [31] does not consider a functional Mahalanobis semi-distance based on the structural properties of the data, as proposed here.

2.4 Empirical results

This section illustrates the performance of the functional classification procedures presented in Section 2.3 through several Monte Carlo simulations and the analysis of two real datasets.

2.4.1 Monte Carlo Study

The Monte Carlo study considers four different scenarios. The first scenario consists of two Brownian motions defined in the closed interval $I = [0, 1]$ with added means $\mu_1(t) = 20t^{1.1}(1-t)$ and $\mu_2(t) = 20t(1-t)^{1.1}$, respectively. Thus, the first scenario considers two Gaussian processes with different means but a common covariance operator with eigenfunctions $\psi_k(t) = \sqrt{2}\sin((k-0.5)\pi t)$ and associated eigenvalues $\lambda_k = 1/(\pi(k-0.5))^2$, $k = 1, 2, \ldots$ The second scenario is similar to the first one but the second Brownian motion is multiplied by $\sqrt{2}$. Thus, the covariance operators of both processes have the same eigenfunctions but the eigenvalues corresponding to the second process are twice those corresponding to the first process. Finally, the remaining two scenarios are similar to the first two, but they replace the Brownian motions with standardized exponential processes with rate 1 and with the same mean functions and covariance operators as the processes in scenarios 1 and 2, respectively.

Subsequently, 500 datasets are generated composed of $n_1 = 150$ functions from the first process and $n_2 = 125$ functions from the second process. The generated functions

are observed at $J = 100$ equidistant points in the closed interval $I = [0, 1]$. Gaussian errors with mean 0 and variance 0.01 are added to each generated point. Once a dataset is generated, the sample is split into a training sample composed of $n_{10} = 100$ functions of the first process and $n_{20} = 75$ functions of the second process, respectively, and a test sample composed of $n_{11} = 50$ functions of the first process and $n_{21} = 50$ functions of the second process, respectively. The discrete trajectories are converted to functional observations using a B-spline basis of order 6 with 20 basis functions which are enough to fit the data well. Figure 2.1 shows four datasets, once smoothing has been performed, corresponding to the four situations considered. As can be seen, all four appear to be complicated scenarios for classification purposes. Note that we have chosen populations of curves with similar characteristics in order to check the potential of the procedures. In particular, we have chosen scenarios 2 and 4, which are close to scenarios 1 and 3 to show that the methods based on the functional Mahalanobis semi-distances work quite well when the processes have very similar covariance structures. Similarly, we have chosen scenarios 3 and 4, close to 1 and 2, to show that similar results are obtained when non Gaussian scenarios are considered.

We analyze the following: (1) the kNN procedure with seven different functional distances, i.e., the L_1, L_2 and L_∞ distances, given by $d_1(\chi_1, \chi_2) = \int_T |\chi_1(t) - \chi_2(t)|\, dt$, $d_2(\chi_1, \chi_2) = \left(\int_T (\chi_1(t) - \chi_2(t))^2\, dt\right)^{1/2}$, and, $d_\infty(\chi_1, \chi_2) = \sup\{|\chi_1(t) - \chi_2(t)| : t \in T\}$, respectively, where χ_1 and χ_2 are two functional observations, denoted by $kL1$, $kL2$ and kLI, respectively, the functional principal components (FPC) semi-distance assuming either a common or a different covariance operator, denoted by kPC and kPD, respectively, and the functional Mahalanobis (FM) semi-distance assuming either a common or a different covariance operator as proposed in Section 2.3, denoted by kMC and kMD, respectively; (2) the centroid procedure with eight different functional distances, the first seven as in the kNN procedure, denoted by $cL1$, $cL2$, cLI, cPC, cPD, cMC and cMD, respectively, and the distance proposed by Delaigle and Hall [15] given in (2.3.6) and denoted by cDH; (3) the linear and quadratic Bayes classification rules as proposed in Section 2.3, denoted by FLB and FQB, respectively; and (4) the multivariate linear and quadratic Bayes classification rules applied

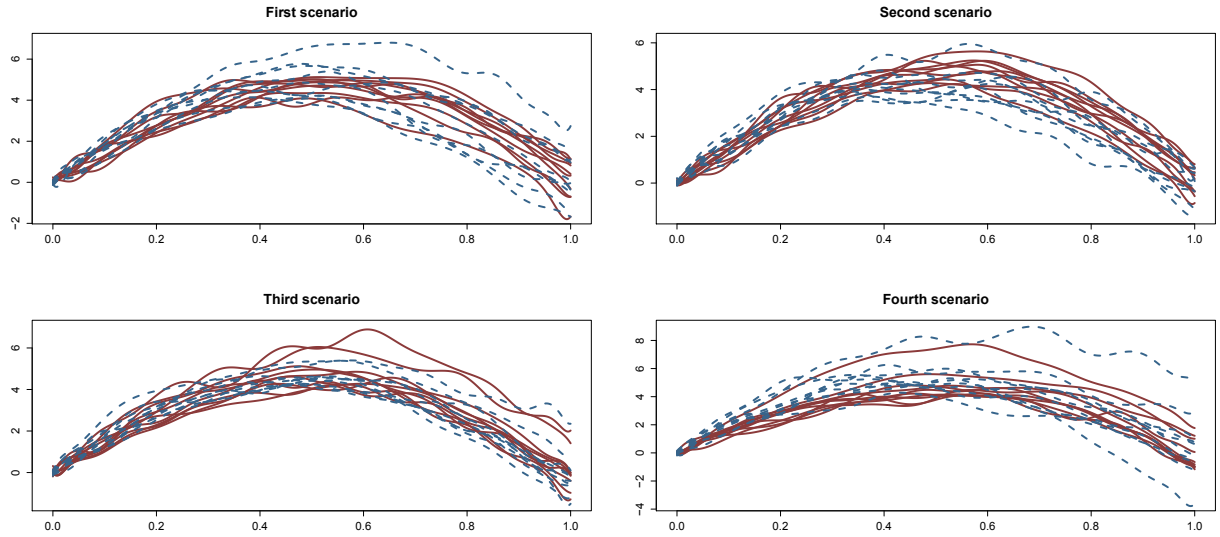


Figure 2.1: B-spline basis approximations of datasets corresponding to the four experiments considered. There are 10 functions of the first process (solid) and another 10 functions of the second process (dashed).

on the coefficients of the B-spline basis representation, denoted by LBC and QBC, respectively. This method can be seen as a simplification of the one proposed by James and Hastie [31]. The values of the threshold parameter K needed to compute the kPC, kPD, kMC, kMD, cPC, cPD, cMC, cMD, cDH, FLB and FQB methods and the number of neighbors in the kNN procedures are determined in the simulation study and in the real data examples using leave-one-out cross-validation on the training samples with a maximum number of 12 eigenfunctions and 9 neighbors, respectively. Once these quantities are selected, the observations in the test samples are classified using the estimates based on all the training samples. The prior probabilities for the FLB and FQB methods are given by the training sample proportions.

Table 2.1 show means and standard deviations of the proportion of correct classification of the test samples for the four scenarios. These results are summarized in the boxplots shown in Figure 2.2. The boxplots only show the best performing methods. The excluded methods are $cL1$, $cL2$, $cLInf$, cPC, cPD, and QBC. On the other hand, Figure 2.3 shows the boxplots of the optimal number of principal components needed

to compute the kPC, kPD, kMC, kMD, cPC, cPD, cMC, cMD, cDH, FLB and FQB methods for the four scenarios. Apparently, there is no a clear pattern about the number of principal components needed. In view of table 2.1 and figure 2.2, we point out the following findings. First, in most of the cases, either the kMC method or the cMC method attains the largest proportion of correct classifications. Second, the proportions of correct classifications for the third and fourth scenarios are slightly larger than the corresponding proportions for the first and second scenarios suggesting that Gaussianity is not necessarily an advantage for classification methods based on the functional Mahalanobis semi-distance. Third, in all the scenarios, classification methods based on the functional Mahalanobis semi-distance have better performance than any other functional distance or semi-distance or any other alternative method such as the one based on the basis functions coefficients. Fourth, at least in these scenarios, the use of the kPD, kMD, cPD and cMD methods is not of practical advantage. Indeed, even if the generated processes have different covariance operators, the methods appear to work better assuming a common covariance operator. However, remember that we are considering two covariance operators that are quite close. Fifth, note that in most of the situations, the spread of good classification rates linked to methods based on the functional Mahalanobis semi-distance is smaller than using any other alternative. In summary, this limited simulation analysis appears to confirm that the functional Mahalanobis semi-distance may be a useful tool for classifying functional observations.

2.4.2 Real data study: Tecator dataset

Next, we apply the classification procedures to the Tecator dataset previously analyzed by Ferraty and Vieu [19], Rossi and Villa [52], Li and Yu [36], Alonso et al. [1] and Martín-Barragán et al. [41], among others. The dataset that consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer is available at http://lib.stat.cmu.edu/datasets/tecator. The absorbance of a meat sample is a function given by $log_{10}(I_0/I)$ where I_0 and I are, respectively, the intensity of the light before and after passing through the meat sample. Each observation consists of a 100-channel absorbance spectrum in the wavelength range $850 - 1050$ nm, contents of

Scenario	kL1	kL2	kLI	kPC	kPD	kMC	kMD	cL1	cL2	cLI	cPC	cPD	cMC	cMD	cDH	FLB	FQB	LBC	QBC
First	.752 (.040)	.754 (.040)	.756 (.041)	.767 (.039)	.767 (.039)	.806 (.035)	.793 (.035)	.674 (.068)	.684 (.066)	.687 (.062)	.688 (.066)	.690 (.066)	**.824** (.038)	.806 (.036)	.800 (.043)	.822 (.037)	.802 (.037)	.816 (.040)	.722 (.047)
Second	.721 (.039)	.722 (.038)	.711 (.039)	.746 (.036)	.747 (.036)	.773 (.033)	.751 (.037)	.647 (.069)	.654 (.068)	.657 (.066)	.658 (.068)	.661 (.068)	**.788** (.038)	.709 (.044)	.763 (.045)	**.788** (.038)	.744 (.041)	.770 (.039)	.617 (.048)
Third	.839 (.037)	.845 (.036)	.844 (.036)	.863 (.034)	.862 (.034)	**.897** (.027)	.878 (.031)	.689 (.073)	.718 (.075)	.725 (.063)	.721 (.075)	.724 (.075)	.837 (.036)	.811 (.040)	.816 (.043)	.821 (.038)	.791 (.042)	.830 (.038)	.714 (.051)
Fourth	.767 (.037)	.769 (.036)	.761 (.036)	.796 (.033)	.796 (.034)	**.812** (.033)	.793 (.037)	.655 (.069)	.680 (.073)	.692 (.066)	.684 (.074)	.687 (.075)	.788 (.039)	.755 (.039)	.761 (.044)	.773 (.040)	.760 (.038)	.784 (.040)	.644 (.052)

Tables 2.1: Means and standard deviations of the proportion of correct classification of the test samples for the four scenarios. The best proportion of correct classification in each scenario is shown in bold.

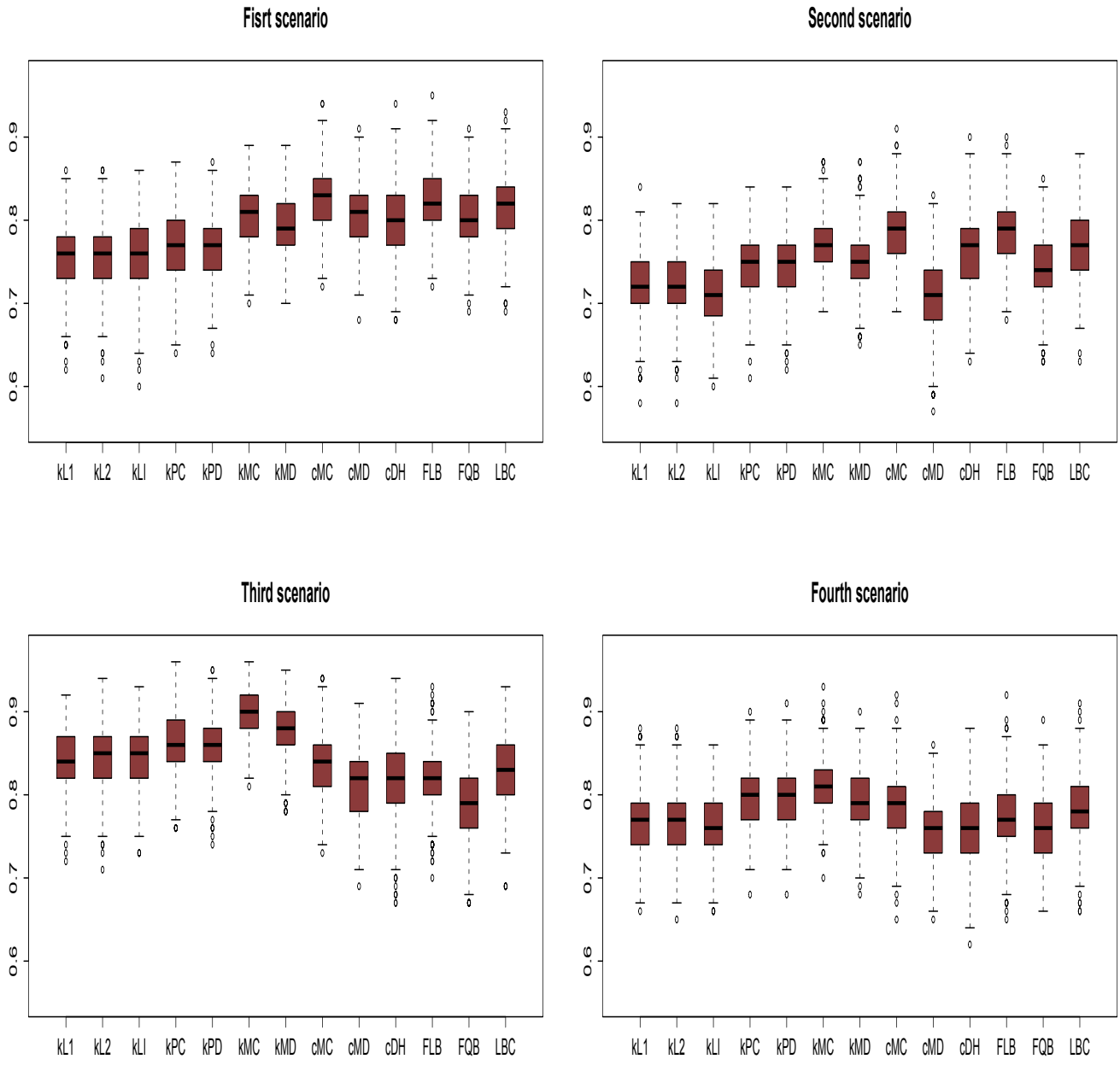


Figure 2.2: Proportions of correct classification for all the scenarios.

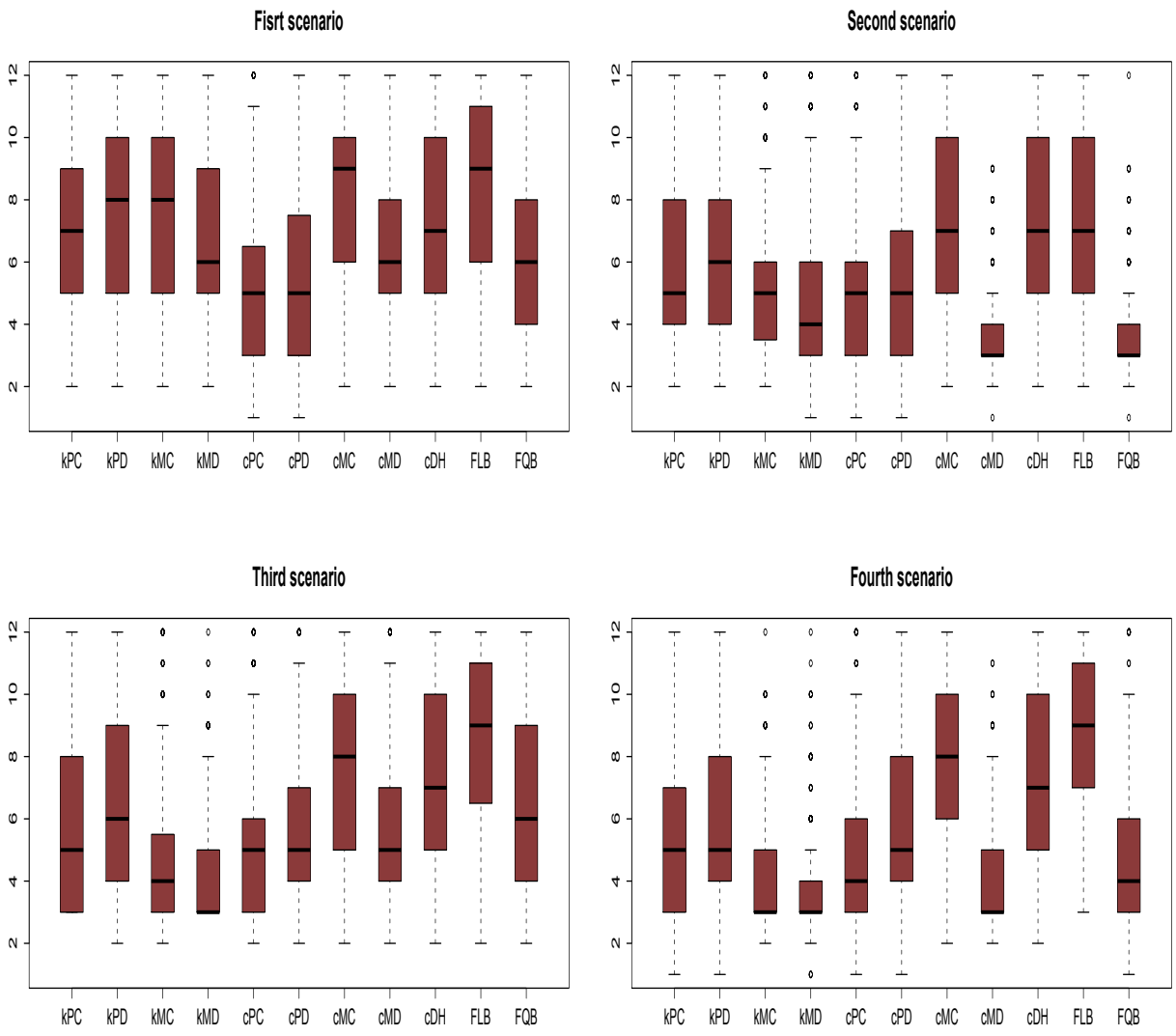


Figure 2.3: Optimal number of principal components for all scenarios.

moisture (water), fat and protein. The recorded absorbance can be seen as a discretized version of the continuous process. The classification problem here is to separate meat samples with a high fat content (more than 20%) from samples with a low fat content (less than 20%) based on absorbance. Among the 215 samples, 77 have high fat content and 138 have low fat content. Previous analyses of this dataset have suggested that classification of the second order derivatives of the observed functions produces lower misclassification rates. Therefore, the analysis of the original data and their second order derivatives are carried out. In both cases, the original discrete observations and their second differences are converted to functional observations using a B-spline basis of order 6 with 20 and 40 basis functions, respectively. Figure 2.4 shows the sample of this 100-channel absorbance spectrum and its second derivatives after smoothing.

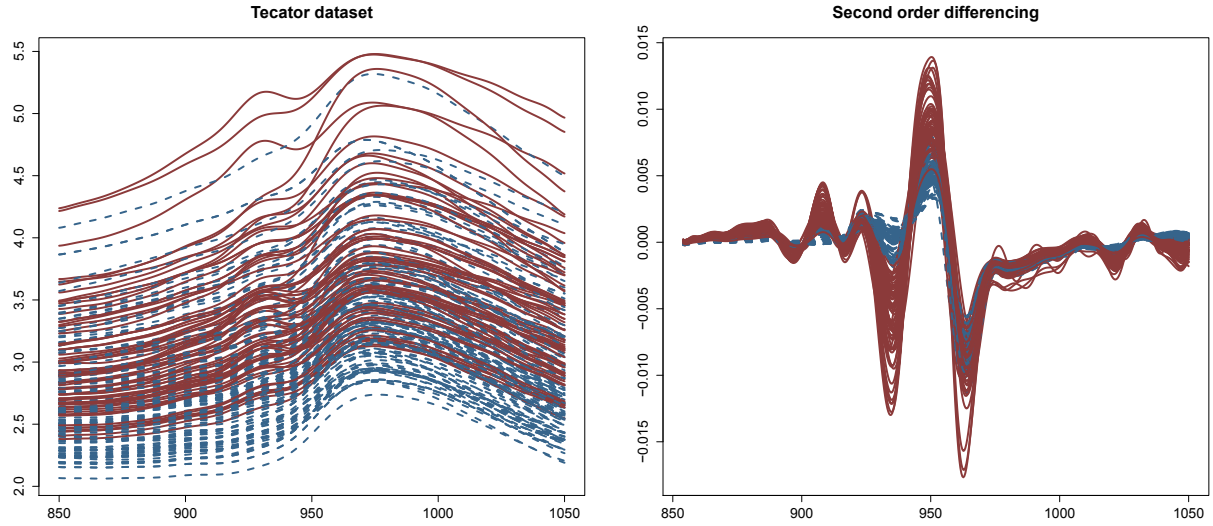


Figure 2.4: Left: Original observations of the Tecator dataset. Right: Second order derivatives. Curves with high fat content in solid lines and with low fat content in dashed lines.

Next, 500 training samples composed of 58 and 104 randomly chosen functions of meat with high fat content and low fat content, respectively, are considered. For each training sample, the remaining 19 and 34 functions with high and low fat content are used as a test sample. Table 2.2 show means and standard deviations of the proportion

of correct classification of the test samples for the two cases. The results of this table are summarized in the boxplots shown in Figure 2.5. The boxplots only show the best performing methods. Interestingly, the methods not shown are the same as the simulations. We use the same range of values on the vertical axes in the two boxplots to make comparisons between the two boxplots easier. In both cases, the kMC method is the winner. The highest mean proportions of correct classification for the Tecator dataset and the second order derivatives with kMC are 0.9834 and 0.9908, respectively, suggesting that it is not necessary to use the second order derivatives of the Tecator data to obtain almost perfect classification. Note that using a similar experiment, Rossi and Villa [52] obtained classification rates of 0.9672 and 0.9740 for the original and second order derivatives with SVMs, respectively; Li and Yu [36] obtained classification rates of 0.9602 and 0.9891 for the original and second order derivatives, respectively, with a segmentation approach; Alonso et al. [1] obtained a classification rate of 0.9798 with a method that takes into account the original, the first and the second order derivatives; and, finally, Martín-Barragán et al. [41] obtained a classification rate of 0.9891 for the second order derivatives with a method based on interpretable SVM classifiers for functional data which have high classification accuracy. Note that all of the previous approaches are more sophisticated than the ones taken here. Finally, boxplots of the optimal number of principal components needed to compute the kPC, kPD, kMC, kMD, cPC, cPD, cMC, cMD, cDH, FLB and FQB methods for the original dataset and their second order derivatives are found in Figure 2.6. Apparently, there is no general rule regarding the number of principal components used.

2.4.3 Real data study: Phoneme dataset

Finally, the classification procedures are applied to the Phoneme dataset described in Ferraty and Vieu [20] and available at http://www.math.univ-toulouse.fr/staph/npfda/npfda-datasets.html. This dataset is a part of the original one analyzed in Hastie, Buja and Tibshirani [28] and available at http://statweb.stanford.edu/∼tibs/ElemStatLearn/datasets/phoneme.data. The dataset contains log-periodograms corresponding to 32-ms recordings of a sample of

Dataset	kL1	kL2	kLI	kPC	kPD	kMC	kMD	cL1	cL2	cLI	cPC	cPD	cMC	cMD	cDH	FLB	FQB	LBC	QBC
Original	.788	.811	.862	.815	.814	**.983**	.971	.680	.683	.698	.684	.684	.964	.951	.949	.953	.965	.926	.891
	(.039)	(.039)	(.037)	(.040)	(.040)	(.010)	(.016)	(.038)	(.037)	(.037)	(.037)	(.037)	(.015)	(.024)	(.032)	(.017)	(.020)	(.024)	(.034)
Sec. Dif.	.989	.986	.982	.990	.988	**.991**	.967	.964	.962	.955	.966	.963	.968	.938	.964	.953	.955	.921	.720
	(.009)	(.009)	(.011)	(.007)	(.009)	(.008)	(.016)	(.018)	(.019)	(.020)	(.017)	(.019)	(.016)	(.026)	(.019)	(.017)	(.019)	(.026)	(.058)

Tables 2.2: Means and standard deviations of the proportion of correct classification of the test samples for the tecator dataset and for their second order differences. The best proportion of correct classification in each case is shown in bold.

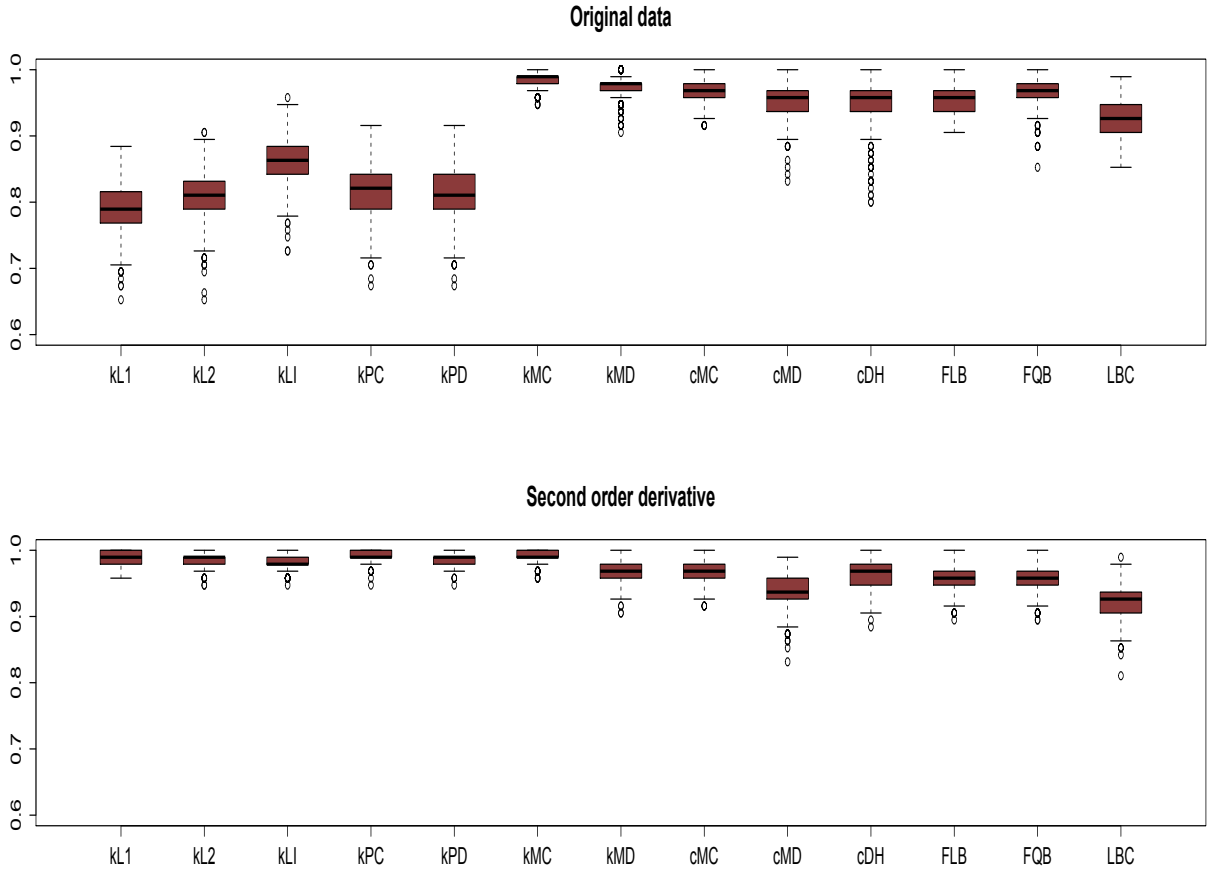


Figure 2.5: Proportion of correct classification for the Tecator dataset. Top: original data. Bottom: second order derivative of the dataset.

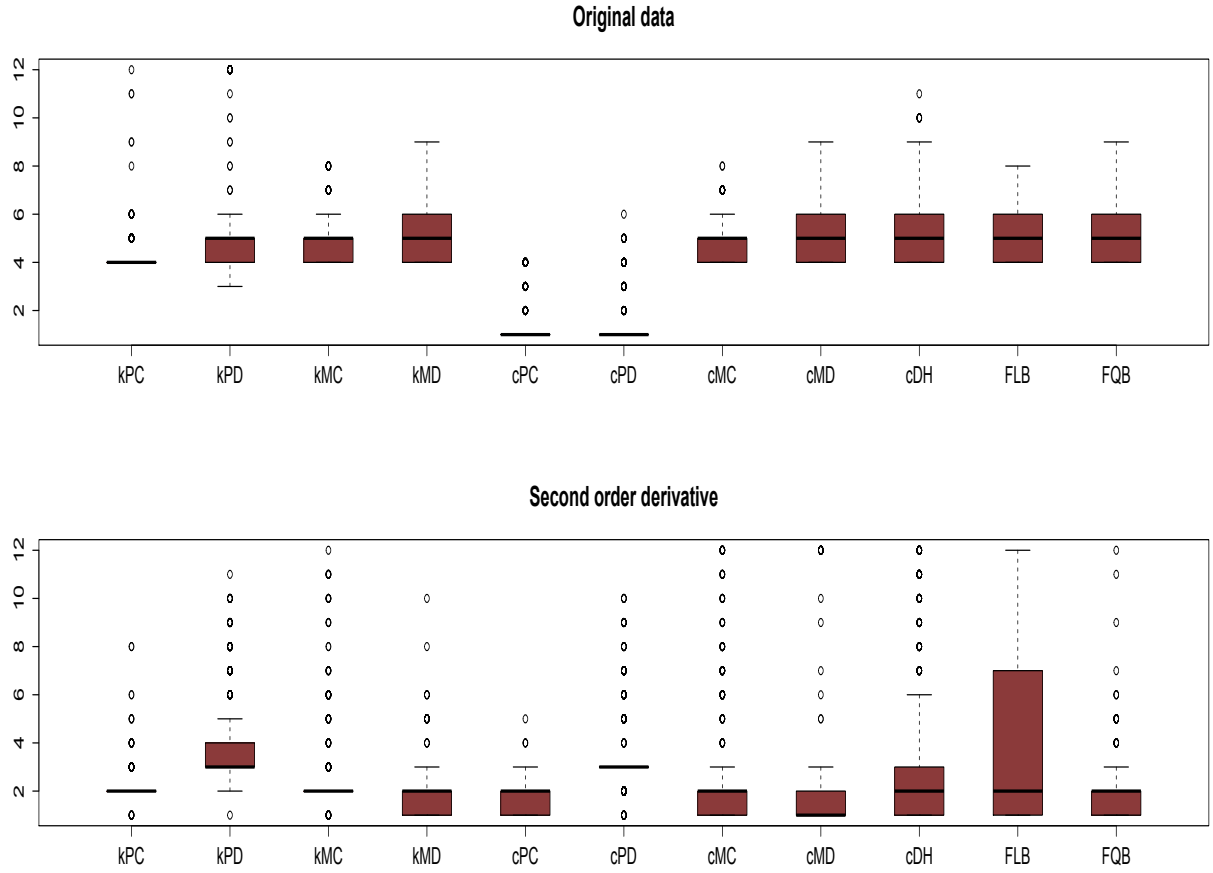


Figure 2.6: Optimal number of principal components for the Tecator dataset. Top: original data. Bottom: second order derivative of the dataset.

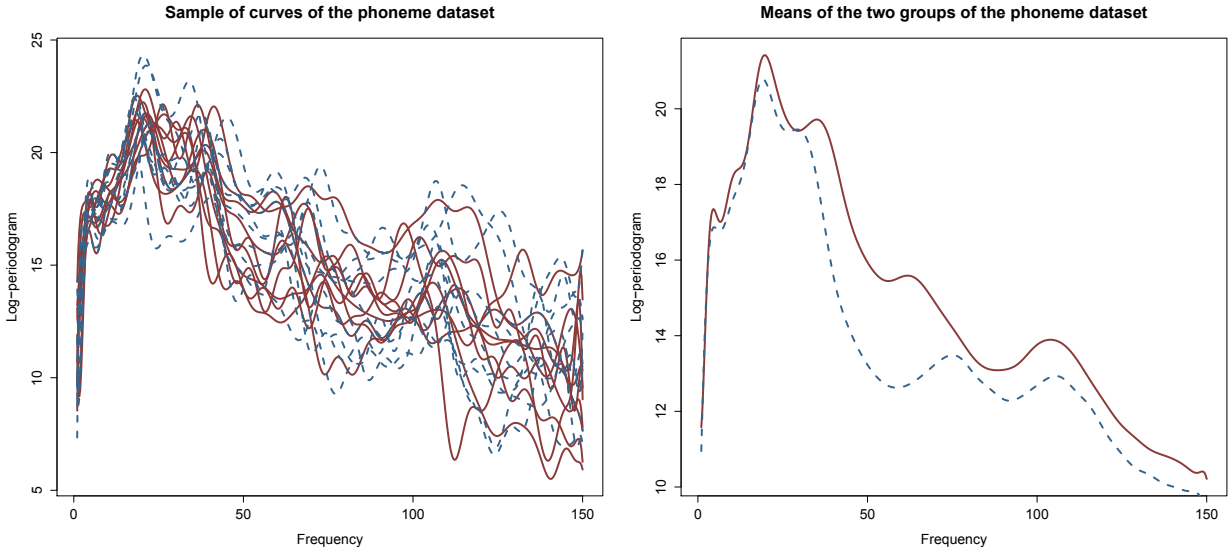


Figure 2.7: Left: Sample of 20 curves of the phoneme dataset (log-periodograms for "aa" in solid lines and log-periodograms for "ao" in dashed lines). Right: Means of the two groups (for "aa" the solid line and log-periodograms for "ao" the dashed line).

speakers. Here, two populations corresponding to the phonemes "aa" as in the vowel of "dark" and "ao" as in the first vowel of "water" are considered. Each speech frame is represented by 400 samples at a 16-kHz sampling rate where only the first 150 frequencies from each subject are retained. The data consists of 800 log-periodograms of length 150, with known class phoneme membership. The classification problem here is to separate the two phonemes. The discrete observations are converted to functional observations using a B-spline basis of order 6 with 40 basis functions, respectively. Figure 2.7 shows a sample of 10 log-periodograms of each class and the means of the two classes for the whole dataset. The figure confirms that it is difficult to distinguish the log-periodograms from one another.

Next, 500 training datasets, each containing 600 curves, are considered. They are composed of 300 randomly chosen log-periodograms of both vowels. For each training sample, the remaining 200 curves (100 per class) are used as a test sample. The means and standard deviations of the proportion of correct classification of the test samples are shown in Table 2.3. These results are summarized in the boxplots shown in Figure

2.8. The boxplots do not show the worst performing methods, which are again the same as in the simulations and the tecator dataset. In this case, the cMC and FLB methods perform the best. Note that with equal class sizes (as is the case here), these two methods are equivalent. The highest mean proportion of correct classification for the Phoneme dataset is 0.8216, which is slightly larger than other alternatives. Boxplots of the optimal number of principal components needed for various methods are found in Figure 2.9. The mean number of functional principal components used with cMC/FLB is around 9. Other methods with worse performance also have mean values close to 9. However, the spreads of the number of principal components used by other methods are larger than with cMC/FLB.

Dataset	kL1	kL2	kLI	kPC	kPD	kMC	kMD	cL1	cL2	cLI	cPC	cPD	cMC	cMD	cDH	FLB	FQB	LBC	QBC
Original	.787 (.022)	.781 (.023)	.781 (.023)	.798 (.021)	.777 (.022)	.808 (.021)	.788 (.020)	.751 (.029)	.736 (.028)	.702 (.029)	.737 (.028)	.733 (.028)	**.822** (.021)	.797 (.021)	.808 (.027)	**.822** (.021)	.797 (.021)	.811 (.024)	.776 (.025)

Tables 2.3: Means and standard deviations of the proportion of correct classification of the test samples for the phoneme dataset. The best proportion of correct classification is shown in bold.

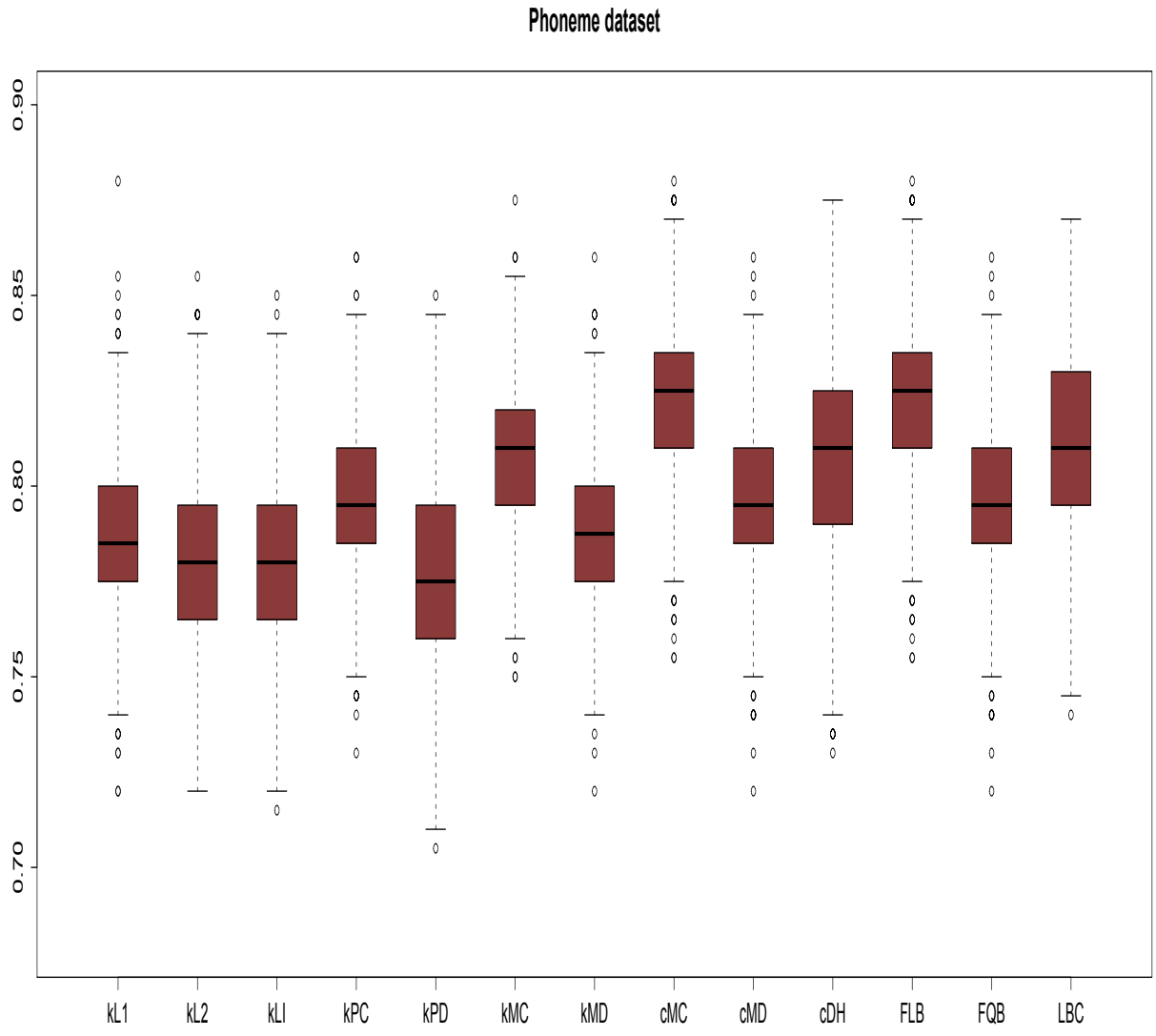


Figure 2.8: Proportion of correct classification for the Phoneme dataset.

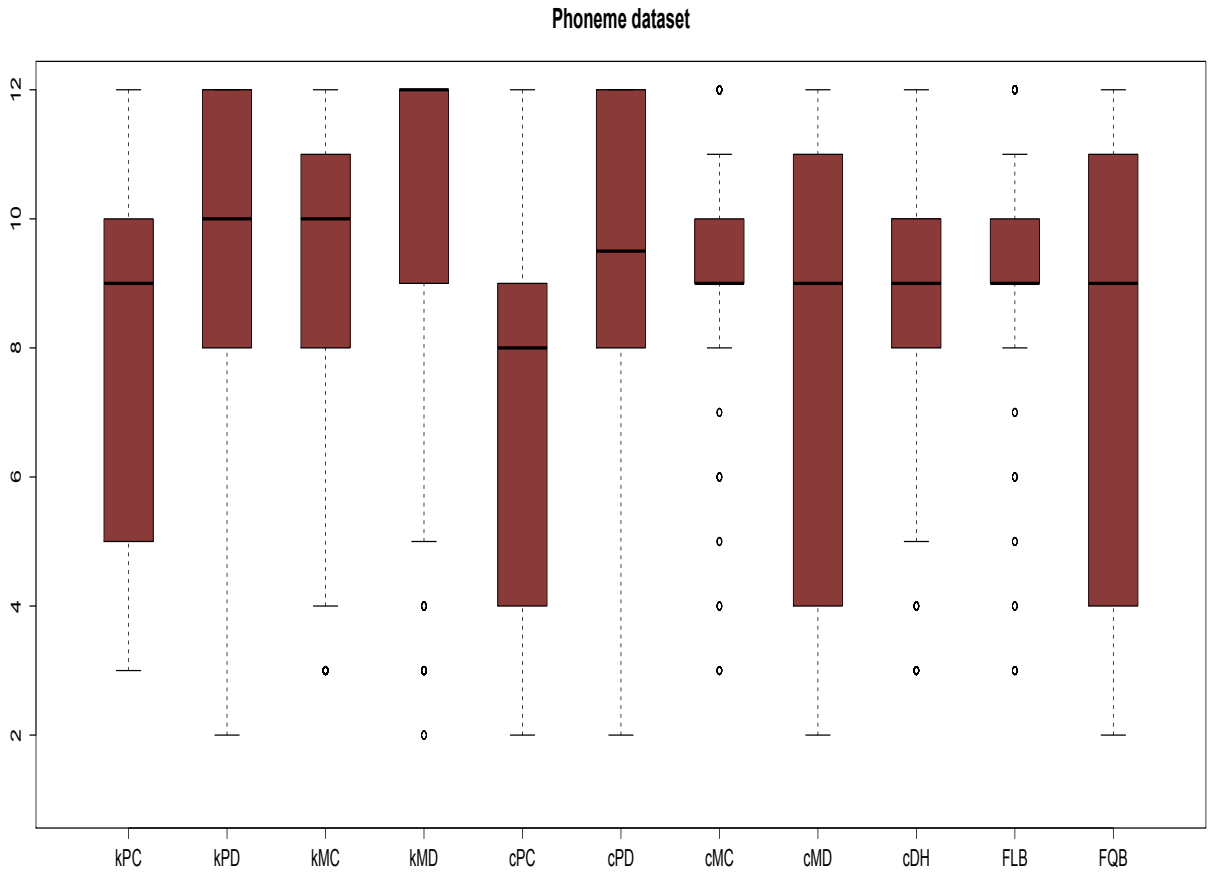


Figure 2.9: Optimal number of principal components for the Phoneme dataset.

2.5 Conclusions

In this chapter, we have introduced a new semi-distance for functional data that generalize the multivariate Mahalanobis distance to the functional framework. We use the regularized square root inverse operator given in Mas [42] which allows to write the functional Mahalanobis semi-distance between an observation and the sample mean function of the set of functions in terms of the standardize functional principal component scores. New versions of several classification procedures including kNN, the centroid method and functional Bayes classification rules have been proposed based on the functional Mahalanobis semi-distance. Monte Carlo experiments and the analysis of two real data examples illustrate the good behavior of the classification methods based on the functional Mahalanobis semi-distance.

CHAPTER 3

Two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance

3.1 Introduction

In the multivariate context, the two-sample Hotelling's T^2 statistic is frequently used to test the equality of means of two independent Gaussian random samples with the same covariance matrix. If equality of covariance matrices is not assumed, the testing issue is known as the multivariate Behrens-Fisher problem although the two-sample Hotelling's T^2 statistic is still used. The common point of the two statistics, that is, assuming that the covariance matrices are equal or that they are different, is that the two-sample Hotelling's T^2 statistics are just the squared Mahalanobis distance between the sample means of both random samples.

Some approaches have been proposed so far to test whether the mean functions of two functional samples are equal. For instance, Fan and Lin [18] developed tests for comparing the means of two functional samples based on the adaptive Neyman test and wavelet thresholding techniques. Horváth and Kokoszka [29] presented procedures for testing the equality of the means in two independent functional random samples

based on the functional principal components semi-distance between the sample means of the two functional samples. The asymptotic distribution of the statistic derived in this way converges, under the null hypothesis, to weighted sums of squares of independent standard Gaussians. As alternative and trying to avoid the use of the weighted asymptotic distribution, Horváth and Kokoszka [29] also proposed a normalized version of the statistic based on the functional principal components semi-distance that has a chi-square limit. These inferential procedures were extended to the case of functional time series in Horváth et al. [30].

In this Chapter, we focus on the problem of testing the equality of mean functions in two random samples independently drawn from two functional distributions. Specially, we derive two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance assuming either a common or a different covariance operator for the random samples following the ideas developed in the multivariate context. We show that the test statistics derived in terms of the Mahalanobis semi-distance coincide with the normalized test statistic proposed by Horváth and Kokoszka [29], although, these authors did not consider the functional Mahalanobis semi-distance in the development of their normalized statistic. Therefore, we establish the link between the Hotelling's T^2 statistic in the multivariate and functional settings. Finally, we will illustrate with the scenarios previously considered in Section 2.4.1 and a real data example from climatology the performance of the test statistics based on the functional Mahalanobis semi-distance and the functional principal components semi-distance.

This chapter is organized as follows. Section 3.2 introduces some preliminaries needed to define properly the statistics associated to the homogeneity test. Section 3.3 introduces the statistics based on the functional Mahalanobis semi-distance for testing the equality of mean functions in two independent random samples and describes their asymptotic behavior. Sections 3.4 and 3.5 evaluate the performance of the procedures proposed in Section 3.3 by means of a simulation study and a real data application. Finally, some conclusions are drawn in Section 3.6.

3.2 Preliminaries

The aim of this section is to briefly review the multivariate Hotelling's T^2 statistics to motivate their extension to the functional framework.

3.2.1 Multivariate Hotelling's T^2 statistics

Let $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ and $\mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ be two random samples independently drawn from two multivariate Gaussian distributions with means $\mathbf{m}_{\mathbf{x}_1}$ and $\mathbf{m}_{\mathbf{x}_2}$ and positive definite covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively. The aim is to test:

$$H_0 : \mathbf{m}_{\mathbf{x}_1} = \mathbf{m}_{\mathbf{x}_2} \qquad vs. \qquad H_A : \mathbf{m}_{\mathbf{x}_1} \neq \mathbf{m}_{\mathbf{x}_2}. \tag{3.2.1}$$

Let $\widehat{\mathbf{m}}_{\mathbf{x}_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{x}_{1i}$ and $\widehat{\mathbf{m}}_{\mathbf{x}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$ be the sample means of the two random samples, respectively. The Multivariate Hotelling's T^2 statistic for the test (3.2.1) is given by:

$$T^2 = d_M\left(\widehat{\mathbf{m}}_{\mathbf{x}_1}, \widehat{\mathbf{m}}_{\mathbf{x}_2}\right)^2, \tag{3.2.2}$$

where $d_M\left(\widehat{\mathbf{m}}_{\mathbf{x}_1}, \widehat{\mathbf{m}}_{\mathbf{x}_2}\right)$ is the Mahalanobis distance between $\widehat{\mathbf{m}}_{\mathbf{x}_1}$ and $\widehat{\mathbf{m}}_{\mathbf{x}_2}$ defined as:

$$d_M\left(\widehat{\mathbf{m}}_{\mathbf{x}_1}, \widehat{\mathbf{m}}_{\mathbf{x}_2}\right)^2 = (\widehat{\mathbf{m}}_{\mathbf{x}_1} - \widehat{\mathbf{m}}_{\mathbf{x}_2})'\widehat{\mathbf{C}}_{12}^{-1}(\widehat{\mathbf{m}}_{\mathbf{x}_1} - \widehat{\mathbf{m}}_{\mathbf{x}_2}), \tag{3.2.3}$$

and $\widehat{\mathbf{C}}_{12}$ is an estimate of the covariance matrix of $\widehat{\mathbf{m}}_{\mathbf{x}_1} - \widehat{\mathbf{m}}_{\mathbf{x}_2}$ defined depending on whether $\mathbf{C}_1$ and $\mathbf{C}_2$ are assumed to be equal or not. Hence, if $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, the covariance matrix of $\widehat{\mathbf{m}}_{\mathbf{x}_1} - \widehat{\mathbf{m}}_{\mathbf{x}_2}$ is given by:

$$\mathbf{C}_{12} = \frac{n_1 + n_2}{n_1 n_2}\mathbf{C},$$

that can be estimated with:

$$\widehat{\mathbf{C}}_{12} = \frac{n_1 + n_2}{n_1 n_2}\widehat{\mathbf{C}}, \tag{3.2.4}$$

where $\widehat{\mathbf{C}}$ in (3.2.4) is the pooled covariance matrix given by:

$$\widehat{\mathbf{C}} = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)\,\widehat{\mathbf{C}}_1 + (n_2 - 1)\,\widehat{\mathbf{C}}_2\right),$$

and $\widehat{\mathbf{C}}_1$ and $\widehat{\mathbf{C}}_2$ are the sample covariance matrices of $\mathbf{C}_1$ and $\mathbf{C}_2$ based on the two random samples, respectively, given by:

$$\widehat{\mathbf{C}}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\mathbf{x}_{ji} - \widehat{\mathbf{m}}_{\mathbf{x}_j})(\mathbf{x}_{ji} - \widehat{\mathbf{m}}_{\mathbf{x}_j})', \tag{3.2.5}$$

for $j = 1, 2$, respectively. Let T_C^2 denotes the multivariate Hotelling's T^2 statistic in (3.2.2) where $\widehat{\mathbf{C}}_{12}$ is given in (3.2.4). Then, $\frac{n-p-1}{p(n-2)}T_C^2$ follows a F distribution with p and $n - p - 1$ degrees of freedom under the null hypothesis of equality of means given in (3.2.1). The previous F distribution asymptotically tends to the χ_p^2 distribution. On the other hand, if $\mathbf{C}_1 \neq \mathbf{C}_2$, the covariance matrix of $\widehat{\mathbf{m}}_{\mathbf{x}_1} - \widehat{\mathbf{m}}_{\mathbf{x}_2}$ is given by:

$$\mathbf{C}_{12} = \frac{1}{n_1}\mathbf{C}_1 + \frac{1}{n_2}\mathbf{C}_2,$$

that can be estimated through:

$$\widehat{\mathbf{C}}_{12} = \frac{1}{n_1}\widehat{\mathbf{C}}_1 + \frac{1}{n_2}\widehat{\mathbf{C}}_2, \tag{3.2.6}$$

where $\widehat{\mathbf{C}}_1$ and $\widehat{\mathbf{C}}_2$ are defined in (3.2.5). Let T_D^2 denotes the multivariate Hotelling's T^2 statistic in (3.2.2) where $\widehat{\mathbf{C}}_{12}$ is given in (3.2.6). Then, the distribution of T_D^2 under the null hypothesis in (3.2.1) has been approximated with several scaled F distributions, see James [32], Yao [61], Johansen [33], Nel and van der Merwe [45] and Kim [34], among others.

3.3 Functional two-sample Hotelling's T^2 statistics

The purpose of this section is to introduce the functional two-sample Hotelling's T^2 statistics defined through the functional Mahalanobis semi-distance proposed by Galeano

et al. [23]. For that, we adapt the definitions of the two-sample Hotelling's T^2 statistics, T_C^2 and T_D^2, for multivariate data defined in Section 3.2

Let χ_1 and χ_2 be two independent functional random variables defined in the infinite dimensional space $L^2(T)$, with mean functions $\mu_{\chi_1}(t) = E[\chi_1(t)]$ and $\mu_{\chi_2(t)} = E[\chi_2(t)]$ and compact covariance operators Γ_{χ_1} and Γ_{χ_2}, respectively. Therefore, χ_1 and χ_2 can be written as $\chi_1 = \mu_{\chi_1} + \epsilon_1$ and $\chi_2 = \mu_{\chi_2} + \epsilon_2$, respectively, where ϵ_1 and ϵ_2 are two independent error functional random variables defined in $L^2(T)$ with compact covariance operators Γ_{χ_1} and Γ_{χ_2}, respectively. Additionally, we assume that $E\left[\|\epsilon_1\|_2^4\right] < \infty$ and $E\left[\|\epsilon_2\|_2^4\right] < \infty$, respectively.

Let $\chi_{11}, \ldots, \chi_{1n_1}$ and $\chi_{21}, \ldots, \chi_{2n_2}$ be two random samples independently drawn from χ_1 and χ_2, respectively, that satisfies:

$$\chi_{1i}(t) = \mu_{\chi_1}(t) + \epsilon_{1i}(t), \tag{3.3.1}$$

for $1 \leq i \leq n_1$, and

$$\chi_{2i}(t) = \mu_{\chi_2}(t) + \epsilon_{2i}(t), \tag{3.3.2}$$

for $1 \leq i \leq n_2$, respectively, where $\epsilon_{11}, \ldots, \epsilon_{1n_1}$ and $\epsilon_{21}, \ldots, \epsilon_{2n_2}$ are two random samples independently drawn from ϵ_1 and ϵ_2, respectively. The aim is to test:

$$H_0 : \mu_{\chi_1} = \mu_{\chi_2} \qquad vs. \qquad H_A : \mu_{\chi_1} \neq \mu_{\chi_2}. \tag{3.3.3}$$

Let $\widehat{\mu}_{\chi_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\chi_{1i}$ and $\widehat{\mu}_{\chi_2} = \frac{1}{n_2}\sum_{i=1}^{n_2}\chi_{2i}$ be the sample mean functions of the two random samples, respectively, and let Γ_{12}, be the covariance operator of $\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}$. Similarly as in (3.2.2), we propose to test the equality of means using the functional Hotelling's T^2 statistic given by:

$$T_F^2 = d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2, \tag{3.3.4}$$

where $d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})$ is the functional Mahalanobis semi-distance between $\widehat{\mu}_{\chi_1}$ and $\widehat{\mu}_{\chi_2}$ defined as:

$$d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 = \left\langle \widehat{\Gamma}_{K,12}^{-1/2}(\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}), \widehat{\Gamma}_{K,12}^{-1/2}(\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}) \right\rangle. \tag{3.3.5}$$

where $\widehat{\Gamma}_{K,12}^{-1/2}$ is an estimate of the regularized squared root inverse covariance operator of $\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}$ given in (2.2.1). The estimate $\widehat{\Gamma}_{K,12}^{-1/2}$ is defined depending on whether Γ_{χ_1} and Γ_{χ_2} are assumed to be equal or not. In both cases, as shown in the Appendix, the functional Mahalanobis semi-distance in (3.3.5), and thus, the test statistic T_F^2, can be expressed as follows:

$$d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 = \sum_{k=1}^{K} \frac{\widehat{\theta}_{12k}^2}{\widehat{\lambda}_k}, \tag{3.3.6}$$

where $\widehat{\theta}_{12k} = \left\langle \widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}, \widehat{\psi}_k \right\rangle$, for $k = 1, 2, \ldots$ are the functional principal component scores with $\widehat{\psi}_1, \ldots, \widehat{\psi}_K$ and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_K$ being the eigenfunctions and associated eigenvalues, respectively, of $\widehat{\Gamma}_{12}$, an estimate of Γ_{12}, that will be given below. Consequently, the functional Hotelling's T^2 statistic T_F^2 in (3.3.4), can be written using the expression in (3.3.6) that, as mentioned before, depends on whether Γ_{χ_1} and Γ_{χ_2} are assumed to be equal or not.

On the one hand, if $\Gamma_{\chi_1} = \Gamma_{\chi_2} = \Gamma_{\chi}$, the covariance operator of $\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}$, is given by:

$$\Gamma_{12} = \frac{n_1 + n_2}{n_1 n_2} \Gamma_{\chi},$$

that can be estimated with:

$$\widehat{\Gamma}_{12} = \frac{n_1 + n_2}{n_1 n_2} \widehat{\Gamma}_{\chi}, \tag{3.3.7}$$

where $\widehat{\Gamma}_{\chi}$ is the pooled covariance operator given by:

$$\widehat{\Gamma}_{\chi}(\eta) = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)\, \widehat{\Gamma}_{\chi_1}(\eta) + (n_2 - 1)\, \widehat{\Gamma}_{\chi_2}(\eta) \right),$$

for $\eta \in L^2(T)$, and $\widehat{\Gamma}_{\chi_1}$ and $\widehat{\Gamma}_{\chi_2}$ being the sample covariance operators of Γ_{χ_1} and Γ_{χ_2} based on the two random samples, respectively, given by:

$$\widehat{\Gamma}_{\chi_j}(\eta) = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left\langle \chi_{ji} - \widehat{\mu}_{\chi_j}, \eta \right\rangle (\chi_{ji} - \widehat{\mu}_{\chi_j}), \tag{3.3.8}$$

for $j = 1, 2$, respectively. Now, eigenfunctions of $\widehat{\Gamma}_{12}$ are those of $\widehat{\Gamma}_{\chi}$, while the associated eigenvalues are $\frac{n_1+n_2}{n_1 n_2}$ times those of $\widehat{\Gamma}_{\chi}$. The statistic (3.3.6) derived in this way is the functional Hotelling's T^2 statistic assuming a common covariance operator for both samples and will be denoted by T^2_{FC}.

On the other hand, if $\Gamma_{\chi_1} \neq \Gamma_{\chi_2}$, the covariance operator of $\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}$ is given by:

$$\Gamma_{12} = \frac{1}{n_1}\Gamma_{\chi_1} + \frac{1}{n_2}\Gamma_{\chi_2},$$

that can be estimated through:

$$\widehat{\Gamma}_{12} = \frac{1}{n_1}\widehat{\Gamma}_{\chi_1} + \frac{1}{n_2}\widehat{\Gamma}_{\chi_2}, \tag{3.3.9}$$

where $\widehat{\Gamma}_{\chi_1}$ and $\widehat{\Gamma}_{\chi_2}$ are given in (3.3.8). Nevertheless, (3.3.9) is not the empirical covariance operator of a functional sample, as occurs in the previous case. Thus, eigenfunctions and eigenvalues of $\widehat{\Gamma}_{12}$ in (3.3.9) cannot be computed from a data set built in terms of the initial data set. For that reason, we will use the following bootstrap procedure to estimate eigenfunctions and eigenvalues of $\widehat{\Gamma}_{12}$:

Step 1 Let $b = 1$.

Step 2 Obtain a random sample with replacement from $\chi_{11}, \ldots, \chi_{1n_1}$ and another one from $\chi_{21}, \ldots, \chi_{2n_2}$. Denote both bootstrap samples by $\chi^b_{11}, \ldots, \chi^b_{1n_1}$ and $\chi^b_{21}, \ldots, \chi^b_{2n_2}$, respectively.

Step 3 Obtain the functional sample means of the bootstrap samples, denoted by $\widehat{\mu}^b_{\chi_1}$ and $\widehat{\mu}^b_{\chi_2}$, respectively and their difference $\widehat{\mu}^b_{12} = \widehat{\mu}^b_{\chi_1} - \widehat{\mu}^b_{\chi_2}$.

Step 4 Repeat Steps 2 and 3 B times to obtain B bootstrap samples $\widehat{\mu}_{12}^b$, for $b = 1, \ldots, B$. Then, the covariance operator Γ_{12} is estimated with the sample covariance operator of $\widehat{\mu}_{12}^1, \ldots, \widehat{\mu}_{12}^B$, from which we obtain the set of estimated eigenfunctions and associated eigenvalues needed to compute (3.3.6).

The statistic (3.3.6) derived in this way is the functional Hotelling's T^2 statistic assuming different covariance operators for both samples and will be denoted by T_{FD}^2.

To analyze the convergence of the statistics T_{FC}^2 and T_{FD}^2 under the null and alternative hypotheses, we briefly review the statistics given in Horváth and Kokoszka [29] to test (3.3.3). Firstly, Horváth and Kokoszka [29] proposed to use the statistic based on the L^2 distance defined as:

$$U = \frac{n_1 n_2}{n_1 + n_2} d_2(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 = \frac{n_1 n_2}{n_1 + n_2} \int_T \left(\widehat{\mu}_{\chi_1}(t) - \widehat{\mu}_{\chi_2}(t)\right)^2 dt. \tag{3.3.10}$$

Under the conditions given at the beginning of this section and assuming that

$$\frac{n_1}{n_1 + n_2} \to \nu$$

with some $0 \leq \nu \leq 1$, the asymptotic distribution of (3.3.10) under the null hypothesis is the distribution of $\sum_{k=1}^{\infty} \tau_k z_k^2$, where $\tau_1 \geq \tau_2 \geq \ldots$ denotes the eigenvalues of the operator $(1-\nu)\,\Gamma_{\chi_1} + \nu \Gamma_{\chi_2}$ and z_k are independent standard Gaussian random variables. As these eigenvalues are unknown, alternatively, Horváth and Kokoszka [29] considered the statistic:

$$U_F = \frac{n_1 n_2}{n_1 + n_2} d_{PC}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 = \sum_{k=1}^{K} \widehat{\theta}_{12k}^2, \tag{3.3.11}$$

where d_{PC}^K denoted the functional principal components semi-distance introduced in Ferraty and Vieu [20] and $\widehat{\theta}_{121} \geq \widehat{\theta}_{122} \geq \ldots$ are the functional principal component scores of $\widehat{\Gamma}_{12}$. In other words, the idea is to replace in (3.3.10) the L^2 distance with the functional principal components semi-distance. The asymptotic distribution of the

statistic U_F in (3.3.11) under the null hypothesis is the distribution of $\sum_{k=1}^{K} \tau_k z_k^2$, for which critical values can be obtained by simulation. Nevertheless, to avoid the use of simulation, Horváth and Kokoszka [29] proposed a normalized version of U_F given by:

$$NU_F = \sum_{k=1}^{K} \frac{\widehat{\theta}_{12k}^2}{\widehat{\lambda}_k},$$

that has an asymptotic χ_K^2 distribution, see Theorem 5.3 in Horváth and Kokoszka [29]. Now, the statistic NU_F is just the functional Hotelling's T^2 statistic in (3.3.6) that, consequently, inherits the χ_K^2 asymptotic distribution. Additionally, Theorem 5.4 in Horváth and Kokoszka [29] establishes the consistency of the NU_F statistic to reject the null hypothesis if the means are different. For that, it is necessary to assume that $\mu_{\chi_1} - \mu_{\chi_2}$ is not orthogonal to the linear span of $\psi_1, \ldots, \psi_K$. This consistency result is also inherited by the functional Hotelling's T^2 statistics. In the following, we denote by U_{FC} and U_{FD}, the statistic U_F when assuming a common or a different covariance operator of the random samples under analysis.

Finally, the threshold parameter K deserves some comments. In practice, the functional Hotelling's T^2 statistics T_{FC}^2 and T_{FD}^2, as well as the statistics U_{FC} and U_{FD} based on the functional principal components semi-distance, can be used to solve the testing problem with several values of K. Then, one can compare the results of the tests. However, it would be advisable to define a procedure that chooses an appropriate value of K to make a unique decision when this hypothesis test is applied to real data. Galeano et al. [23] propose to select K to compute the functional Mahalanobis semi-distance in classification problems by cross-validation. However, this method can not be easily extended in the hypothesis testing framework. Alternatively, we choose the threshold value K via the cumulative percentage of total variance (CPV), that is the classical approach for determining the number of sample principal components to retain. The cumulative percentage of total variance is defined as follows:

$$CPV\,(k) = \frac{\sum_{j=1}^{k} \widehat{\lambda}_j}{\sum_{j=1}^{k_{\max}} \widehat{\lambda}_j}, \tag{3.3.12}$$

where $\widehat{\lambda}_j$ are the eigenvalues of $\widehat{\Gamma}_{12}$ and $k_{\max}$ is the total number of estimated eigenvalues. The CPV in (3.3.12) is an increasing function that tends to 1. Then, we select the value of K as the value of k from which the function CPV grows very slowly to 1. This is the method that we use in the simulated and real data examples in Sections 3.4 and 3.5.

3.4 Empirical Results

This section illustrates the performance of the test statistics presented in Section 3.3 through several Monte Carlo simulations. In particular, we compare the empirical sizes and powers of the test statistics based on the functional Mahalanobis semi-distance, T^2_{FC} and T^2_{FD}, with those of the test statistics based on the functional principal components semi-distance, U_{FC} and U_{FD}, when the covariance operators of the two random samples are assumed to be equal and when this is not assumed.

3.4.1 Monte Carlo Study

In this Monte Carlo study, we generate functional data sets following the structure described in (3.3.1) and (3.3.2). In particular, we consider the functional means $\mu_{\chi_1}(t) = 20t^{\rho}(1-t)$ and $\mu_{\chi_2}(t) = 20t(1-t)^{\rho}$, respectively, where $\rho = 1, 1.01, 1.02, 1.03, 1.04, 1.05$. Thus, when $\rho = 1$, the null hypothesis holds, which allows us to calculate empirical sizes associated with the test statistics. However, when $\rho \neq 1$, the alternative hypothesis holds allowing the calculation of the corresponding empirical powers. Note also that the larger ρ, the more different are μ_{χ_1} and μ_{χ_2}, as plotted in Figure 3.1. Then, the power is a function of the parameter ρ.

First, we compare the empirical sizes and powers of the testing procedures when the covariance operators of the two random samples are equal. For that, we consider two different scenarios for the error terms. In the first scenario, we have:

$$\epsilon_1(t) = \sum_{k=1}^{\infty} \lambda_k^{1/2} z_{1k} \psi_k(t) \qquad \text{and} \qquad \epsilon_2(t) = \sum_{k=1}^{\infty} \lambda_k^{1/2} z_{2k} \psi_k(t),$$

where $\psi_k(t) = \sqrt{2}\sin((k-0.5)\pi t)$, $t \in [0,1]$, for $k = 1, 2, \ldots$ are the eigenfunctions of the covariance operator of the error functions with associated eigenvalues $\lambda_k = 1/(\pi(k-0.5))^2$, for $k = 1, 2, \ldots$, and z_{1k} and z_{2k} are independent standard Gaussian distributed, for $k = 1, 2, \ldots$ Thus, ϵ_1 and ϵ_2 are two Brownian motions with a common covariance operator. In the second scenario, z_{1k} and z_{2k} are replaced with e_{1k} and e_{2k}, that are independent standardized exponential distributed with rate 1. We consider four configurations of sample sizes (n_1, n_2) given by $(50, 50)$, $(50, 100)$, $(100, 100)$ and $(100, 200)$, respectively. We choose these pairs in order to see how the sample sizes influence the test results.

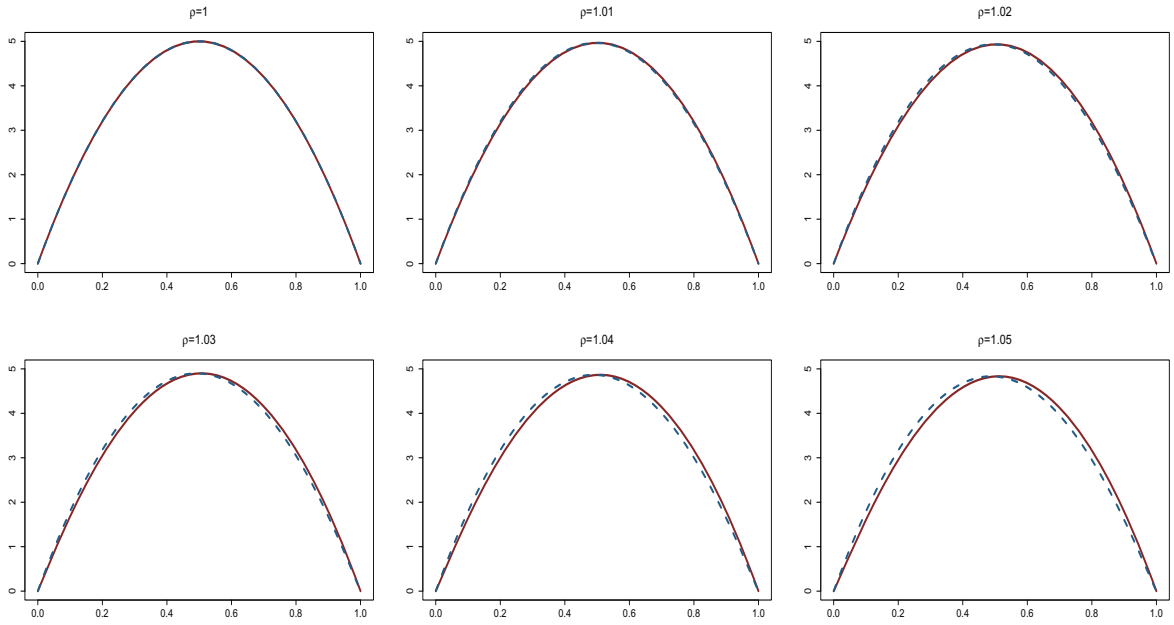


Figure 3.1: Mean functions for different values of ρ. In solid, first sample, and in dashed, second sample.

Subsequently, 1000 data sets are generated of each scenario and pair of sample sizes. The generated functions are observed at $J = 100$ equidistant points in the closed interval $I = [0, 1]$. Gaussian errors with mean 0 and variance 0.01 are added to each generated point. To compute the test statistics, the discrete trajectories are converted to functional observations using a B-spline basis of order 6 with 20 basis functions which seem enough to fit the data well. Figure 3.2 shows a data set generated from the first scenario with

$\rho = 1.05$ and sample size pair $(10, 10)$ with the corresponding sample means. Note that it would be difficult to affirm through visual evaluation that the mean generating functions are different.

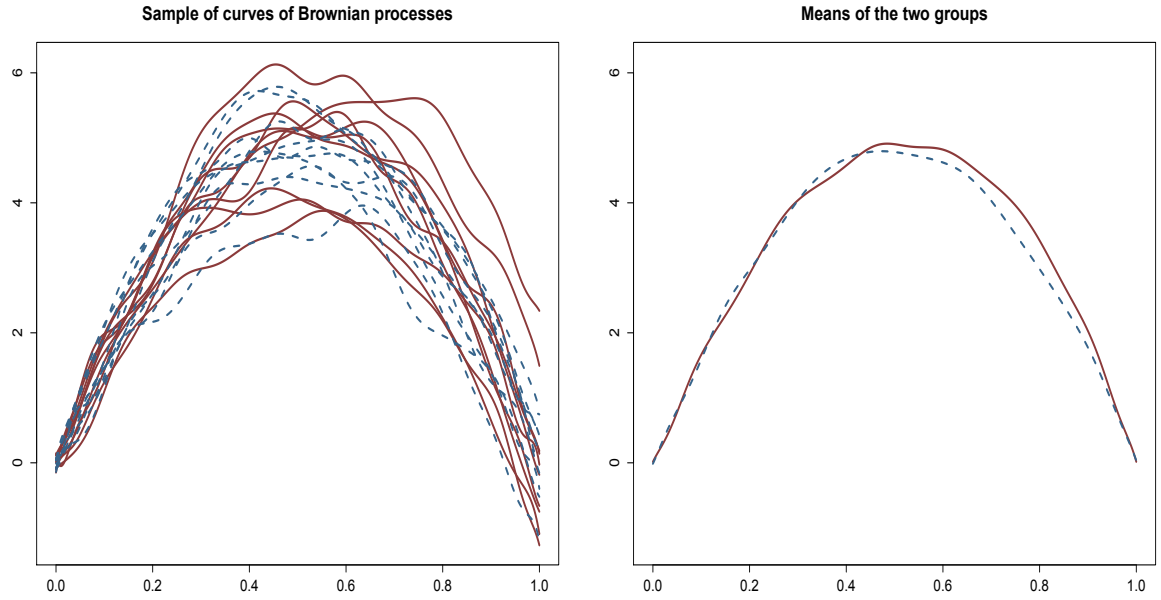


Figure 3.2: Left: Set of 10 functions of the Brownian Motion plus mean $\mu_{\chi_1}(t) = 20t^{1.05}(1-t)$ (solid) and another set of 10 functions of the Brownian Motion plus mean $\mu_{\chi_2}(t) = 20t(1-t)^{1.05}$ (dashed). Right: the sample functional means for the first (solid) and second (dashed) set of curves.

As mentioned in Section 3.3, the functional Hotelling's T^2 statistics can be computed for several values of K, for $K = 1, 2, \ldots$. Nevertheless, it would be preferable to select an appropriate threshold value K and this is done using the CPV criterion in (3.3.12). In this case, we know the true eigenvalues of the covariance operators considered in the Monte Carlo study. These eigenvalues are proportional to those of the covariance operator of the difference of the sample means so that we can use the cumulative percentages obtained from them to select an appropriate threshold K. The first ten cumulative percentages are given by 0.8216, 0.9129, 0.9458, 0.9625, 0.9717, 0.9795, 0.9843, 0.9880, 0.9908 and 0.9931, respectively. As can be seen, the cumulative percentages grow very slowly from the fifth eigenvalue. Thus, we take $CPV = 0.97$ and select the K such that the principal components explain at least the 97% of the variance. After that,

we compute the T^2_{FC} and U_{FC} statistics. Obviously, in practice (with real data) the eigenvalues of the covariance operator of the difference of sample means are unknown. Nevertheless, to take an appropriate CPV, we can use those eigenvalues estimated from the samples using the methods previously described.

The results are summarized in Figure 3.3 and Tables 3.1 and 3.2. On the one hand, Figure 3.3 shows a barplot of the values of K selected by CPV for the 1000 data sets generated for the case of $\rho = 0$ in scenario 1. As can be seen, K only takes values ranging from 4 to 7, being $K = 5$ and $K = 6$ the most frequent values. On the other hand, Tables 3.1 and 3.2 show the empirical sizes and powers of the test statistics for the two scenarios. Each cell in the tables displays the empirical size or power over the 1000 generated data sets. Empirical sizes and powers are calculated at the nominal sizes $\alpha = 0.1, 0.05, 0.01$. In view of these tables, several comments are in order. First, the empirical sizes of the two test statistics are very close to the corresponding nominal sizes in most of the cases. Indeed, the empirical sizes appears to tend to the nominal sizes as the sample sizes increase. Second, if one of the sample sizes is 50, the test statistics have empirical sizes slightly larger than the nominal sizes. Third, in terms of power, the functional Hotelling's T^2 statistic, T^2_{FC}, clearly dominates the test statistic based on the functional principal components semi-distance, U_{FC}, in all the cases. Fourth, the functional Hotelling's T^2 test statistic has good and similar power for both Gaussian and exponential data sets suggesting that non-Gaussianity is not a drawback for T^2_{FC}. Fifth, when the parameter ρ increases, the power of U_{FC} increases slower than that for T^2_{FC}. Sixth, the larger the sample size, the larger the power of both statistics. In summary, we conclude that the functional Hotelling's T^2 statistic appears to outperform the test statistic based on the functional principal components semi-distance in terms of power.

Next, we compare the empirical sizes and powers of T^2_{FD} and U_{FD} when the covariance operators of the two random samples are different. As before, we consider two different scenarios for the error terms. In the first scenario, we have:

$$\epsilon_1(t) = \sum_{k=1}^{\infty} \lambda_{1k}^{1/2} z_{1k} \psi_k(t) \qquad \text{and} \qquad \epsilon_2(t) = \sum_{k=1}^{\infty} \lambda_{2k}^{1/2} z_{2k} \psi_k(t),$$

Tables 3.1: Empirical sizes and powers of the functional Hotelling's T^2 statistic and the test statistic based on the functional principal components semi-distance when $\Gamma_{\chi_1} = \Gamma_{\chi_2}$ for the first scenario.

n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.149	0.112	0.075	0.063	0.018	0.009
		1.01	0.186	0.111	0.109	0.058	0.031	0.015
50	50	1.02	0.351	0.125	0.232	0.071	0.099	0.015
		1.03	0.660	0.179	0.537	0.099	0.316	0.023
		1.04	0.865	0.230	0.802	0.124	0.595	0.035
		1.05	0.973	0.293	0.947	0.154	0.844	0.041
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.129	0.112	0.068	0.056	0.023	0.013
		1.01	0.200	0.105	0.133	0.063	0.036	0.010
50	100	1.02	0.497	0.153	0.378	0.091	0.183	0.019
		1.03	0.783	0.183	0.673	0.097	0.432	0.017
		1.04	0.951	0.283	0.898	0.141	0.757	0.039
		1.05	0.993	0.460	0.980	0.237	0.943	0.058
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.109	0.101	0.054	0.048	0.013	0.007
		1.01	0.217	0.113	0.145	0.049	0.043	0.008
100	100	1.02	0.616	0.156	0.486	0.079	0.266	0.018
		1.03	0.925	0.235	0.872	0.114	0.694	0.036
		1.04	0.995	0.444	0.986	0.248	0.944	0.053
		1.05	1.000	0.678	1.000	0.410	0.998	0.119
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.107	0.101	0.050	0.059	0.005	0.013
		1.01	0.263	0.114	0.172	0.056	0.062	0.012
100	200	1.02	0.709	0.182	0.602	0.101	0.360	0.018
		1.03	0.972	0.292	0.934	0.154	0.842	0.038
		1.04	0.999	0.596	0.997	0.320	0.988	0.079
		1.05	1.000	0.878	1.000	0.616	1.000	0.165

Tables 3.2: Empirical sizes and powers of the functional Hotelling's T^2 statistic and the test statistic based on the functional principal components semi-distance when $\Gamma_{\chi_1} = \Gamma_{\chi_2}$ for the second scenario.

n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.113	0.111	0.065	0.053	0.017	0.019
		1.01	0.193	0.128	0.107	0.066	0.033	0.018
50	50	1.02	0.387	0.139	0.283	0.067	0.113	0.012
		1.03	0.655	0.185	0.533	0.093	0.317	0.016
		1.04	0.868	0.265	0.796	0.135	0.620	0.039
		1.05	0.970	0.347	0.953	0.175	0.850	0.033
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.102	0.111	0.054	0.046	0.014	0.014
		1.01	0.205	0.108	0.127	0.059	0.035	0.013
50	100	1.02	0.480	0.115	0.332	0.056	0.146	0.007
		1.03	0.781	0.193	0.666	0.096	0.447	0.032
		1.04	0.946	0.293	0.900	0.169	0.766	0.043
		1.05	0.995	0.444	0.988	0.237	0.940	0.064
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.098	0.096	0.052	0.051	0.014	0.007
		1.01	0.225	0.118	0.131	0.062	0.047	0.014
100	100	1.02	0.638	0.165	0.514	0.076	0.272	0.022
		1.03	0.913	0.266	0.851	0.137	0.685	0.038
		1.04	0.991	0.459	0.982	0.242	0.945	0.067
		1.05	1.000	0.690	1.000	0.406	0.995	0.123
n_1	n_2	ρ	T^2_{FC} 10%	U_{FC} 10%	T^2_{FC} 5%	U_{FC} 5%	T^2_{FC} 1%	U_{FC} 1%
		1.00	0.100	0.112	0.043	0.051	0.010	0.013
		1.01	0.278	0.116	0.175	0.074	0.078	0.016
100	200	1.02	0.710	0.162	0.590	0.080	0.351	0.017
		1.03	0.957	0.317	0.933	0.180	0.835	0.037
		1.04	1.000	0.566	0.999	0.309	0.989	0.088
		1.05	1.000	0.874	1.000	0.634	1.000	0.188

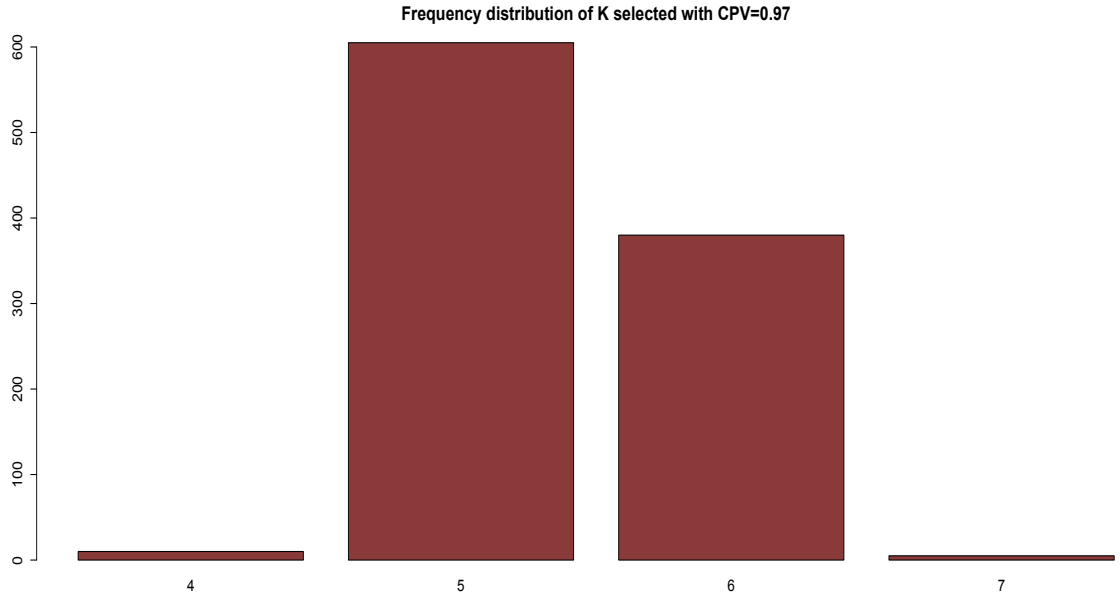


Figure 3.3: Values of K selected in the 1000 simulations with Gaussian processes.

where $\psi_k(t) = \sqrt{2}\sin((k-0.5)\pi t)$, $t \in [0,1]$, for $k = 1, 2, \ldots$ are the eigenfunctions of the covariance operator of the error functions with associated eigenvalues $\lambda_{1k} = 1/(\pi(k-0.5))^2$ and $\lambda_{2k} = 2/(\pi(k-0.5))^2$, for $k = 1, 2, \ldots$, for the first and second random samples, respectively, and z_{1k} and z_{2k} are independent standard Gaussian distributed, for $k = 1, 2, \ldots$ Thus, ϵ_1 and ϵ_2 are two Brownian motions with the same eigenfunctions but with the eigenvalues corresponding to the second error process twice those corresponding to the first error process. In the second scenario, and similarly to the case of common covariance operators, z_{1k} and z_{2k} are replaced with e_{1k} and e_{2k}, that are independent standardized exponential distributed with rate 1.

Then, 1000 data sets are generated of each pair of sample sizes and scenario with the same configurations of samples sizes and generation mechanism as in the first set of simulations. For each generated data set, we obtain $B = 1000$ bootstrap samples as explained in Section 3.3 allowing us to obtain the eigenfunctions and eigenvalues of the estimated covariance operator of the difference of the sample means of the two random samples. Then, in order to fix the value of CPV used in the simulation study to compute

T^2_{FD} and U_{FD}, we compute the mean bootstrap eigenvalues based on the 1000 data sets. A visual inspection of these eigenvalues for each pair of sample sizes and scenario leaded us to select $CPV = 0.97$ in all the situations. Subsequently, once the value of K has been fixed, we compute the two statistics for each generated data set.

The results are summarized in Tables 3.3 and 3.4 that show the empirical sizes and powers of the test statistics for the two scenarios. As in the previous case, each cell in the tables displays the empirical size or power calculated at the nominal sizes $\alpha = 0.1, 0.05, 0.01$ over the 1000 generated data sets. The results in terms of sizes and powers of the simulation study when the covariance operators of the random samples are different are very similar to those when the covariance operators of the random samples are the same. In particular, we would like to note that the bootstrap procedure does not appear to have a significant effect on the limit behavior of the test statistics. Finally, we repeated the study with $B = 10000$ bootstrap replications obtaining similar results, which for brevity are omitted in this Chapter.

3.5 Real data study

In this section, we compare the results obtained by the functional Hotelling's T^2 statistics and the test statistics based on the functional principal components semi-distance with the Canadian Temperature data set previously analyzed by Ramsay and Silverman [48] and Zhang and Chen [63], among others. The data set contains the daily temperature records of 35 Canadian weather stations over a year (365 days). As in Zhang and Chen [63], the 35 stations have been split in three regions, resulting in 15 stations in the Eastern region, another 15 stations in the Western region and the remaining 5 stations in the Northern region. See Table 3.5 to see the stations assigned in each of the three regions. Following Ramsay and Silverman [48] and Ramsay et al. [47], the discrete observations are converted to functional observations using a Fourier series basis with 65 basis functions. Figure 3.4 shows the smoothed temperature curves of the Eastern (solid), Western (dashed) and Northern (dotted) weather stations and the estimated mean temperature functions of these regions. As can be seen, the mean temperature

Tables 3.3: Empirical sizes and powers of the functional Hotelling's T^2 statistic and the test statistic based on the functional principal components semi-distance when $\Gamma_{\chi_1} \neq \Gamma_{\chi_2}$ for the first scenario.

n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.113	0.102	0.066	0.046	0.017	0.012
		1.01	0.156	0.103	0.089	0.058	0.024	0.013
50	50	1.02	0.304	0.118	0.20	0.064	0.084	0.015
		1.03	0.473	0.150	0.358	0.070	0.159	0.019
		1.04	0.713	0.164	0.604	0.094	0.373	0.021
		1.05	0.868	0.246	0.790	0.125	0.598	0.033
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.115	0.122	0.065	0.057	0.019	0.009
		1.01	0.171	0.105	0.083	0.051	0.029	0.009
50	100	1.02	0.366	0.115	0.260	0.059	0.122	0.010
		1.03	0.654	0.169	0.531	0.087	0.308	0.020
		1.04	0.894	0.259	0.820	0.142	0.608	0.033
		1.05	0.977	0.296	0.945	0.155	0.855	0.036
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.117	0.091	0.062	0.052	0.009	0.014
		1.01	0.189	0.099	0.119	0.048	0.039	0.009
100	100	1.02	0.465	0.155	0.358	0.078	0.150	0.016
		1.03	0.753	0.203	0.642	0.102	0.426	0.015
		1.04	0.934	0.289	0.891	0.143	0.761	0.034
		1.05	0.997	0.451	0.992	0.252	0.958	0.057
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.100	0.110	0.052	0.058	0.017	0.013
		1.01	0.238	0.089	0.146	0.042	0.043	0.007
100	200	1.02	0.611	0.160	0.494	0.073	0.274	0.018
		1.03	0.903	0.231	0.853	0.129	0.668	0.035
		1.04	0.991	0.413	0.982	0.223	0.933	0.047
		1.05	1.000	0.688	1.000	0.416	0.994	0.121

Tables 3.4: Empirical sizes and powers of the functional Hotelling's T^2 statistic and the test statistic based on the functional principal components semi-distance when $\Gamma_{\chi_1} \neq \Gamma_{\chi_2}$ for the second scenario.

n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.130	0.120	0.065	0.071	0.017	0.013
		1.01	0.156	0.129	0.092	0.066	0.021	0.019
50	50	1.02	0.273	0.116	0.174	0.069	0.056	0.023
		1.03	0.458	0.152	0.319	0.081	0.151	0.021
		1.04	0.720	0.192	0.596	0.102	0.347	0.032
		1.05	0.895	0.258	0.827	0.143	0.632	0.047
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.118	0.109	0.062	0.053	0.015	0.012
		1.01	0.199	0.125	0.123	0.063	0.039	0.014
50	100	1.02	0.362	0.138	0.252	0.077	0.113	0.017
		1.03	0.659	0.181	0.523	0.096	0.330	0.021
		1.04	0.866	0.228	0.786	0.125	0.612	0.030
		1.05	0.969	0.345	0.943	0.180	0.831	0.049
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.110	0.108	0.062	0.053	0.017	0.010
		1.01	0.159	0.110	0.085	0.063	0.020	0.009
100	100	1.02	0.445	0.158	0.309	0.092	0.143	0.024
		1.03	0.741	0.192	0.638	0.108	0.427	0.019
		1.04	0.943	0.299	0.910	0.169	0.777	0.048
		1.05	0.996	0.476	0.991	0.258	0.955	0.087
n_1	n_2	ρ	T^2_{FD} 10%	U_{FD} 10%	T^2_{FD} 5%	U_{FD} 5%	T^2_{FD} 1%	U_{FD} 1%
		1.00	0.115	0.108	0.060	0.054	0.012	0.020
		1.01	0.245	0.114	0.165	0.057	0.057	0.015
100	200	1.02	0.620	0.173	0.507	0.084	0.286	0.018
		1.03	0.917	0.250	0.862	0.123	0.681	0.032
		1.04	0.991	0.426	0.975	0.225	0.930	0.053
		1.05	1.000	0.710	1.000	0.447	0.995	0.134

functions of the stations in the Eastern and Western regions look like similar and far from the mean temperature function of the Northern weather stations.

Tables 3.5: Classification of the Canadian weather stations.

Eastern	St. Johns	Halifax	Sydney	Yarmouth	Charlottesville
	Fredericton	Scheffervll	Arvida	Bagottville	Quebec
	Sherbrooke	Montreal	Ottawa	Toronto	London
Western	Thunderbay	Winnipeg	The Pas	Churchill	Regina
	Pr. Albert	Uranium City	Edmonton	Calgary	Kamloops
	Vancouver	Victoria	Pr. George	Pr. Rupert	Whitehorse
Northern	Dawson	Yellowknife	Iqaluit	Inuvik	Resolute

Based on the reconstructed temperature curves, the objective is to test if the mean temperature functions of the Eastern and Western weather stations during the whole year are the same. We are also interested in testing if the weather stations in the Eastern and Northern and the Western and Northern regions have, respectively, the same mean temperature functions. Before performing the tests, a task that we have to carry out is to verify whether the covariance operators of the groups can be assumed to be the same, in order to choose the appropriate test statistics. For that, Figures 3.5 and 3.6 show the estimated standard deviations and covariance operators surfaces for the curves in the Eastern, Western and Northern regions, respectively, while Figure 3.7 show the corresponding contour plots of the estimated covariance operators. The figures show different shapes and scales suggesting that the covariance operators of the groups are different. Additionally, Figure 3.8 displays the eigenvalues of each estimated covariance operator that appears to move in quite different scales again leading to similar conclusions. Hence, we use the test statistics when the covariance operators of the random samples are assumed to be different.

Next, we compute the statistics T^2_{FD} and U_{FD} for $K = 1, \ldots, 15$ for the three pairs of regions with 1000 bootstrap replications. Table 3.6 displays the p-values of the two test statistics. As can be seen, both testing procedures lead to essentially the same conclusions, rejecting the equality of mean temperature functions between the Eastern

Figure 3.4: Left: Daily temperature of Canada (Eastern weather stations in solid lines, Western weather stations in dashed lines and Northern weather stations in dotted lines). Right: Estimated mean temperature functions of the Eastern, Western and Northern weather stations.

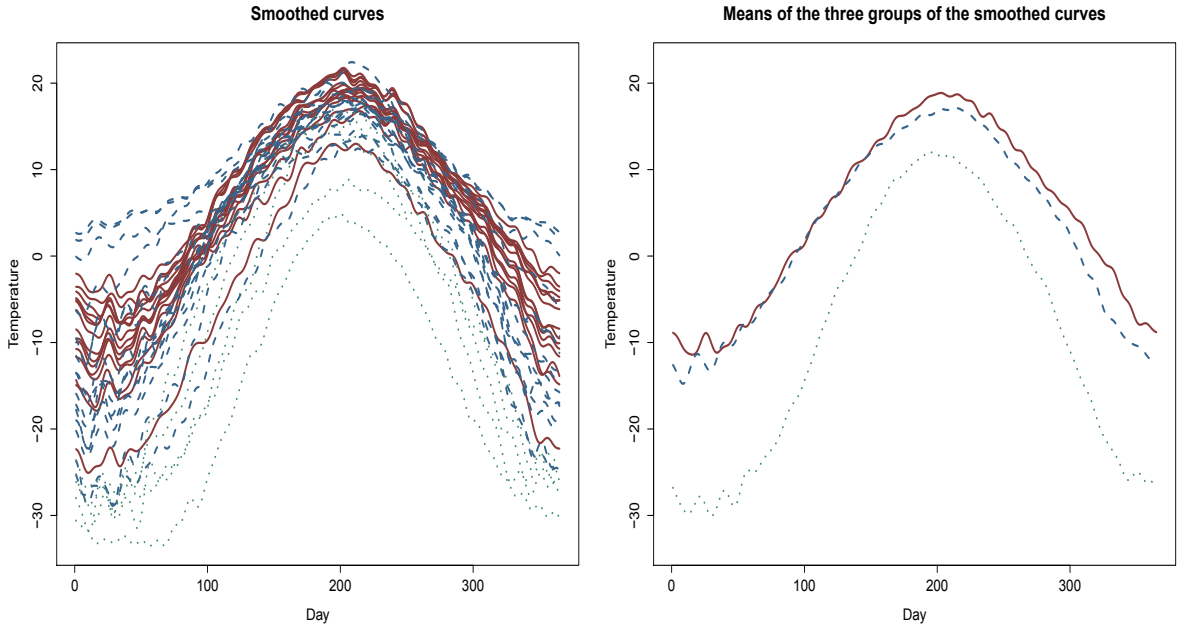
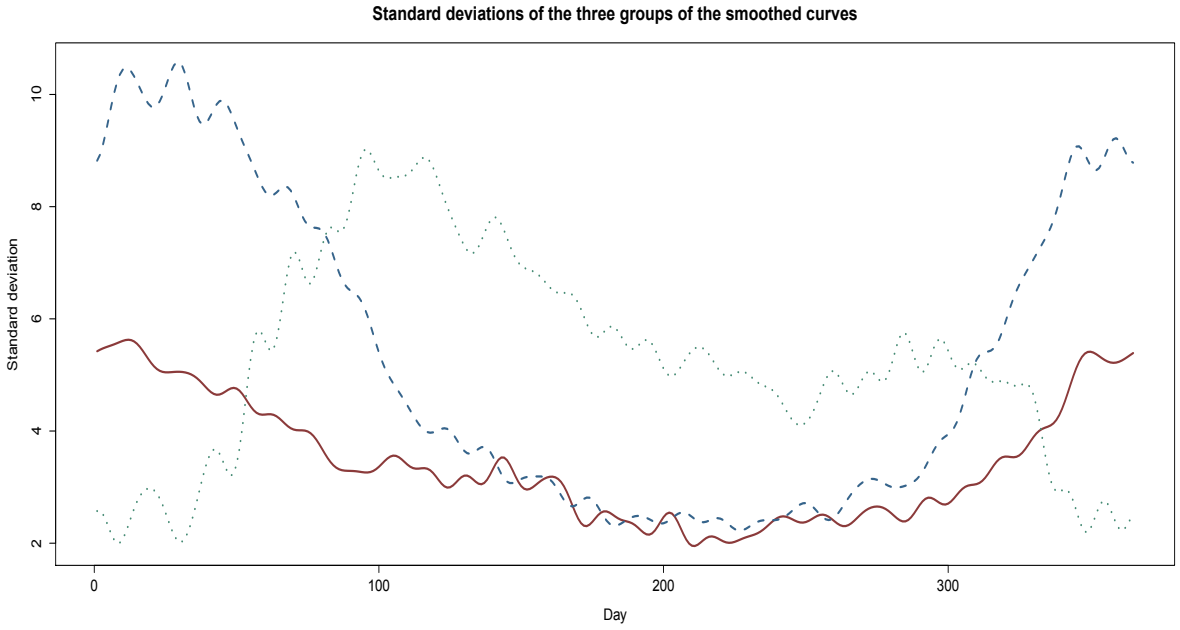


Figure 3.5: Estimated standard deviations of the three groups of the smoothed curves (Eastern weather stations in solid lines, Western weather stations in dashed lines and Northern weather stations in dotted lines).



and Northern regions and Western and Northern regions. However, for Eastern-Western regions, U_{FD} do not reject the null hypothesis of equality of mean functions, while T^2_{FD} reject this null hypothesis when $K > 2$. Then, we select an appropriate value of K using the cumulative percentage of total variance. For that, for each pair of regions, we obtain the eigenvalues of the estimated covariance operator of the difference of the sample means of both random samples obtained as shown in Section 3.3. For the Eastern-Western regions, the cumulative percentage of total variance explained by the first 10 eigenvalues are 0.8749, 0.9787, 0.9925, 0.9947, 0.9962, 0.9972, 0.9977, 0.9983, 0.9987 and 0.9989, for the Eastern-Northern regions, these are given by 0.8822, 0.9452, 0.9755, 0.9960, 0.9981, 0.9992, 0.9995, 0.9997, 0.9998 and 0.9998, while for the Western-Northern pair these are given by 0.7290, 0.9380, 0.9728, 0.9956, 0.9975, 0.9985, 0.9990, 0.9994, 0.9996 and 0.9997. As can be seen, in the three cases, the cumulative percentages grow slowly from 99%. Therefore, we select 99% of the total variation in the three cases. Table 3.6 shows that the value of K selected via the CPV is 3 for Eastern-Western regions and $K = 4$,

Figure 3.6: The estimated covariance operators for the three groups.

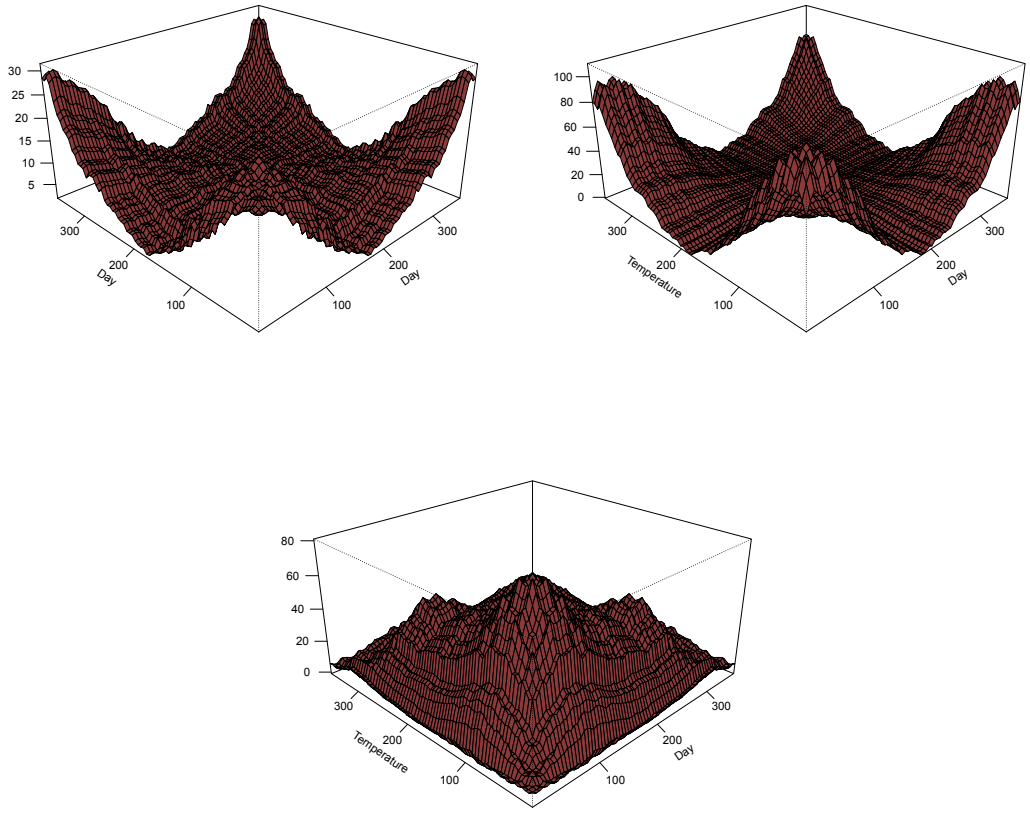


Figure 3.7: The contours of the estimated covariance operators for the three groups.

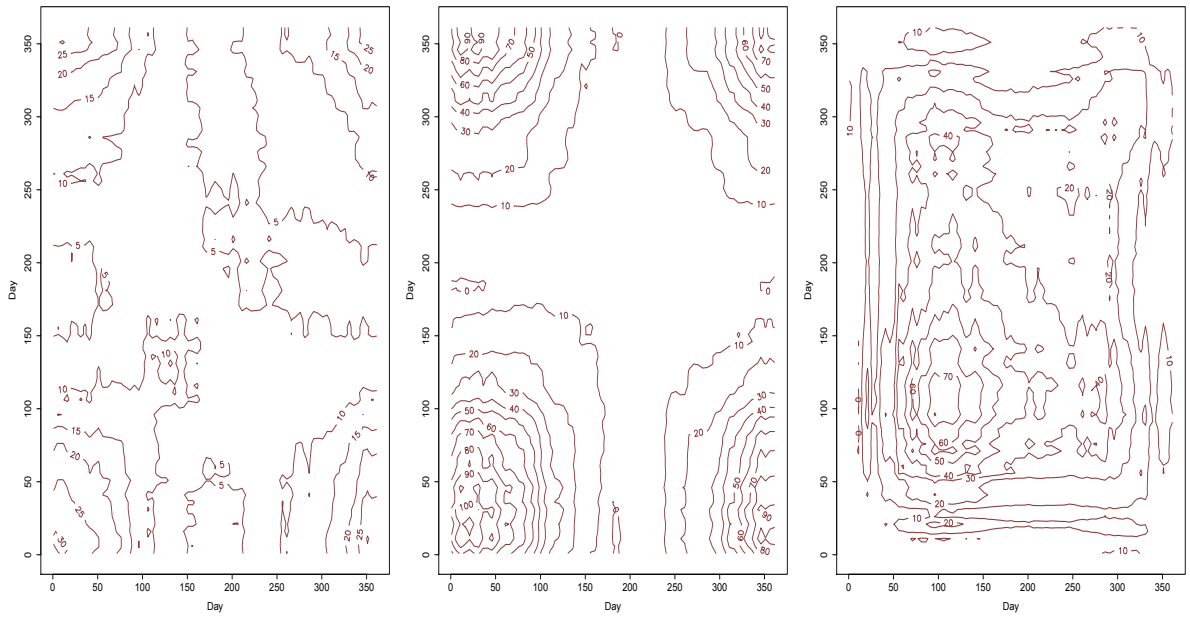
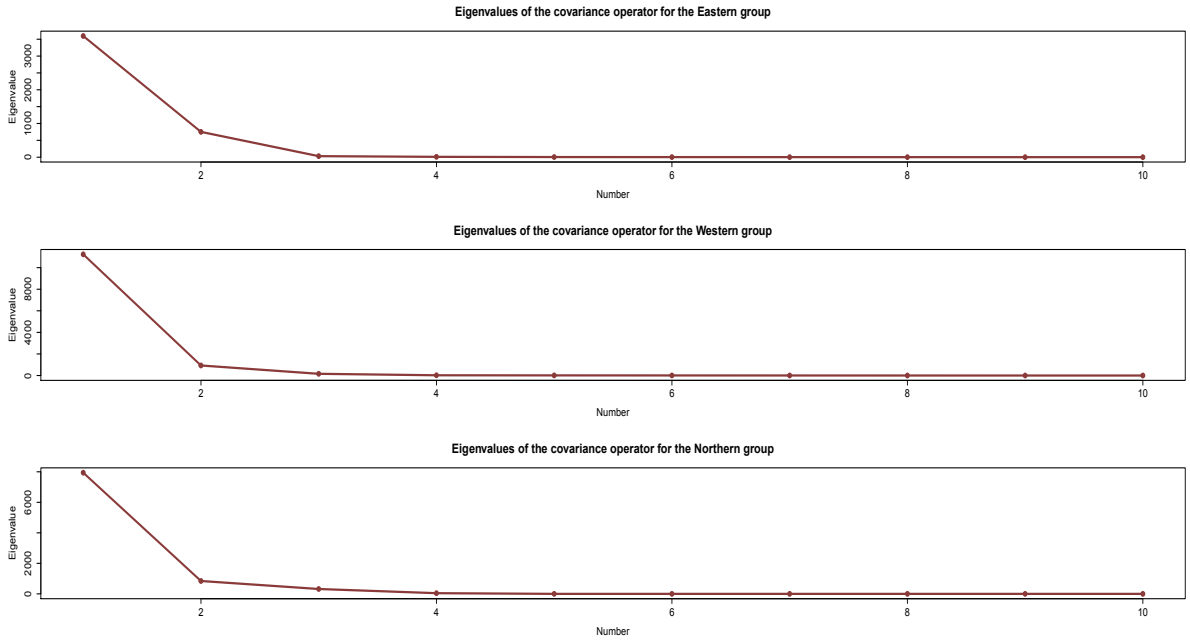


Figure 3.8: The first 10 eigenvalues of the estimated covariance operators for the three groups.



Tables 3.6: P-values (in percent) of the tests based on statistics T^2_{FD} and U_{FD} applied to the Canadian Temperature data set for Eastern-Western, Eastern-Northern and Western-Northern stations.

	Eastern-Western		Eastern-Northern		Western-Northern	
K	T^2_{FD}	U_{FD}	T^2_{FD}	U_{FD}	T^2_{FD}	U_{FD}
1	30.97	31.13	1.17×10^{-5}	0	4.16×10^{-5}	0
2	53.73	34.17	1.10×10^{-22}	0	4.5×10^{-8}	0
3	2.58×10^{-8}	20.63	8.92×10^{-22}	0	2.17×10^{-7}	0
4	7.29×10^{-11}	19.74	1.44×10^{-26}	0	8.01×10^{-7}	0
5	9.14×10^{-15}	19.07	2.55×10^{-30}	0	1.27×10^{-9}	0
6	9.88×10^{-15}	19.52	7.80×10^{-38}	0	9.53×10^{-10}	0
7	7.19×10^{-15}	19.36	3.99×10^{-40}	0	1.00×10^{-9}	0
8	1.56×10^{-15}	19.07	9.88×10^{-139}	0	2.48×10^{-10}	0
9	1.20×10^{-26}	18.79	3.02×10^{-152}	0	1.54×10^{-10}	0
10	1.45×10^{-46}	19.17	2.68×10^{-151}	0	3.96×10^{-10}	0
11	2.32×10^{-91}	18.17	1.41×10^{-181}	0	2.07×10^{-12}	0
12	2.26×10^{-94}	17.62	1.88×10^{-246}	0	4.16×10^{-31}	0
13	9.91×10^{-100}	17.98	5.65×10^{-269}	0	9.13×10^{-36}	0
14	5.57×10^{-99}	18.01	0	0	6.04×10^{-73}	0
15	4.48×10^{-100}	17.52	0	0	1.10×10^{-82}	0
$K - CPV$	2.58×10^{-8}	20.63	1.44×10^{-26}	0	8.01×10^{-7}	0

otherwise. Thus, we conclude that the functional Hotelling's T^2 statistic reject the null hypothesis of equality of mean functions for Eastern-Western regions when K is properly selected.

3.6 Conclusions

In this chapter, we have derived two-sample Hotelling's T^2 statistics for testing the equality of mean functions in two samples independently drawn from two functional distributions based on the functional Mahalanobis semi-distance. In particular, in the case in which the covariance operators of the two random samples are not assumed to be the same, we have proposed a bootstrap method to estimate the covariance operator of the differences between the sample means of the two random samples. The limit distributions of the statistics under the null hypothesis are chi-squared, a result that can be established from the relationship between the proposed statistics and those based on the functional principal components semi-distance proposed in Horváth and Kokoszka [29]. Indeed, we have shown that the derived two-sample Hotelling's T^2 statistics coincide with the normalized functional principal components semi-distance statistics proposed in Horváth and Kokoszka [29]. The simulations and real data application show that the two-sample Hotelling's T^2 statistics appears to outperform the tests based on the functional principal components semi-distance given in Horváth and Kokoszka [29].

3.7 Appendix

Proof of (3.3.6). Using the Fourier decomposition, the difference between the sample functional means $\widehat{\mu}_{\chi_1}$ and $\widehat{\mu}_{\chi_2}$ can be written as:

$$\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2} = \sum_{k=1}^{\infty} \widehat{\theta}_{12k} \widehat{\psi}_k, \tag{3.7.1}$$

where $\widehat{\theta}_{12k} = \left\langle \widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}, \widehat{\psi}_k \right\rangle$ are the scores of $\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}$, for $k = 1, \ldots$ Using the expression (2.2.1), it is straightforward to show that:

$$\begin{aligned} d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 &= \left\langle \widehat{\Gamma}_{K,12}^{-1/2}(\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}), \widehat{\Gamma}_{K,12}^{-1/2}(\widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2}) \right\rangle \\ &= \left\langle \sum_{k=1}^{K} \frac{1}{\widehat{\lambda}_k^{1/2}} \left\langle \widehat{\psi}_k, \widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2} \right\rangle \widehat{\psi}_k, \sum_{k=1}^{K} \frac{1}{\widehat{\lambda}_k^{1/2}} \left\langle \widehat{\psi}_k, \widehat{\mu}_{\chi_1} - \widehat{\mu}_{\chi_2} \right\rangle \widehat{\psi}_k \right\rangle. \end{aligned}$$

Now, from (3.7.1), the previous expression leads to:

$$\begin{aligned} d_{FM}^K(\widehat{\mu}_{\chi_1}, \widehat{\mu}_{\chi_2})^2 &= \left\langle \sum_{k=1}^K \frac{1}{\widehat{\lambda}_k^{1/2}} \left[\left\langle \widehat{\psi}_k, \sum_{j=1}^\infty \widehat{\theta}_{12j} \widehat{\psi}_j \right\rangle \widehat{\psi}_k \right], \sum_{k=1}^K \frac{1}{\widehat{\lambda}_k^{1/2}} \left[\left\langle \widehat{\psi}_k, \sum_{j=1}^\infty \widehat{\theta}_{12j} \widehat{\psi}_j \right\rangle \widehat{\psi}_k \right] \right\rangle \\ &= \left\langle \sum_{k=1}^K \frac{\widehat{\theta}_{12k}}{\widehat{\lambda}_k^{1/2}} \widehat{\psi}_k, \sum_{k=1}^K \frac{\widehat{\theta}_{12k}}{\widehat{\lambda}_k^{1/2}} \widehat{\psi}_k \right\rangle \\ &= \sum_{k=1}^K \frac{\widehat{\theta}_{12k}^2}{\widehat{\lambda}_k} \end{aligned}$$

■

CHAPTER 4

Conclusions

This chapter summarizes the main contributions of the thesis. This dissertation is devoted to functional data analysis, especially to the notion of functional distance. We have proposed a new semi-distance for functional observations, inspired by the Mahalanobis distance frequently used in multivariate data analysis. The functional Mahalanobis semi-distance has been proven useful in supervised classification and hypothesis testing. In the following we present the principal aspects developed in each chapter.

In Chapter 2 we have introduced a new semi-distance for functional data that generalize the multivariate Mahalanobis distance to the functional framework. We use the regularized square root inverse operator given in Mas [42] which allows to write the functional Mahalanobis semi-distance between an observation and the sample mean function of the set of functions in terms of the standardized functional principal component scores. New versions of several classification procedures including kNN, the centroid method and functional Bayes classification rules have been proposed based on the functional Mahalanobis semi-distance. Monte Carlo experiments and the analysis of two real data examples illustrate the good behavior of the classification methods based on the functional Mahalanobis semi-distance.

In Chapter 3 we have derived two-sample Hotelling's T^2 statistics for testing the equality of mean functions in two samples independently drawn from two functional distributions based on the functional Mahalanobis semi-distance. In particular, in the case in which the covariance operators of the two random samples are not assumed to be the same, we have proposed a bootstrap method to estimate the covariance operator of the differences between the sample means of the two random samples. The limit distributions of the statistics under the null hypothesis are chi-squared, a result that can be established from the relationship between the proposed statistics and those based on the functional principal components semi-distance proposed in Horváth and Kokoszka [29]. Indeed, we have shown that the derived two-sample Hotelling's T^2 statistics coincide with the normalized functional principal components semi-distance statistics proposed in Horváth and Kokoszka [29]. The simulations and real data application show that the two-sample Hotelling's T^2 statistics appears to outperform the tests based on the functional principal component semi-distance given in Horváth and Kokoszka [29].

4.1 Research Lines

We close this dissertation presenting some of the issues considered as future research lines:

- In the definitions given in Section 2.2.1, we have assumed that all the functions are aligned on the time axis and hence only differences in amplitude provide information about the distance between two curves. However, functional datasets are often distorted on the time axis. The usual approach to addressing the presence of random variation in time in addition to amplitude variation is to perform time warping. Given the set of curves $\chi_1(t), \ldots, \chi_n(t)$, the idea is to seek a set of time-warping functions $w_i(t)$, for $i = 1, \ldots, n$, such that a new set of functions given by $\widetilde{\chi}_i(t) = \chi_i(w_i(t))$ is well aligned (see Ramsay and Silverman, [48]). Obviously, computing the functional Mahalanobis semi-distance between the original observed functions may not be the most appropriate idea if the data are distorted on the

time axis. There are several possible approaches to this problem. The first is to consider the functional Mahalanobis semi-distance in the dataset of time-warped functions $\widetilde{\chi}_1(t), \ldots, \widetilde{\chi}_n(t)$ rather than the original functions. The second is to consider the functional Mahalanobis semi-distance in the dataset of time-warping functions $w_1(t), \ldots, w_n(t)$. The third is to consider a combination of the two previous approaches. Surely, more research is needed to establish efficient approaches in order to measure distances between functional observations when time warping is performed, and it will be an interesting topic to address in the future.

- To apply the tests in Sections 3.4 and 3.5, it is advisable to select the number of functional principal components used in the computations of the statistics. We propose to use the cumulative percentage of the total variance. However, other selection methods such as the Bayesian information criterion and the Akaike information criterion proposed by Li et al. [37] could be extended to two-sample problems. This would be an objective of future work.

- The range of applications for the new semi-distance is wide and includes clustering and outlier detection for functional data, among others. In the future, it would be interesting to propose new procedures based on the combination of those methods with the functional Mahalanobis semi-distance. In this point, the idea is to adapt methods developed to the multivariate context to the functional scenario. For example, the techniques for outlier detection developed by Filzmoser et al. [21], Maronna and Zamar [40], Becker and Gather [5], Rousseeuw and Van Zomeren [54], Rousseeuw and Van Driessen [53], Rocke and Woodruff [51] and Woodruff and Rocke [60] could be extended to functional framework. As clustering procedures, it will be interesting to extend the works presented by Melnykov and Melnykov [43], Fraley and Raftery [22] and Banfield and Raftery [4]. As mentioned in Section 1.4, the notion of functional distances and semi-distances is useful in various problems including prediction problems, unsupervised classification techniques and for defining density functions for functional variables. In this research line, we propose to extend these functional methodologies proposed by Ferraty and Vieu

[20], Chiou and Li [10] and Delaigle and Hall [14] using the functional Mahalanobis semi-distance. It aims to analyze situations in which the new semi-distance contributes to the improvement of those techniques already developed with alternative semi-distances and distances.

Bibliography

[1] A. M. Alonso, D. Casado, and J. Romo. Supervised classification for functional data: a weighted distance approach. *Computational Statistics and Data Analysis*, 56:2334–2346, 2012.

[2] R. B. Ash and M. F. Gardner. *Topics in stochastic processes*. Academic Press, New York, 1975.

[3] A. Baíllo, A. Cuevas, and J. A. Cuesta-Albertos. Supervised classification for a family of gaussian functional models. *Scandinavian Journal of Statistics*, 38:480–498, 2011.

[4] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[5] C. Becker and U. Gather. The masking breakdown point of multivariate outlier identifications rules. *Journal of the American Statistical Association*, 94(447):947–955, 1999.

[6] M. Benko, W. Härdle, and A. Kneip. Common functional principal components. *The Annals of Statistics*, 37:1–34, 2009.

[7] G. Biau, F. Bunea, and M. H. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[8] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

[9] H. Chen, P.T. Reiss, and T. Tarpey. Otimally Weighted L^2 Distance for Functional Data. *Biometrics*, 70:516–525, 2014.

[10] J.-M. Chiou and P.-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B*, 69:679–699, 2007.

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

[12] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational Statistics and Data Analysis*, 47:111–122, 2004.

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

[14] A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *Annals of Statistics*, 34:1171–1193, 2010.

[15] A. Delaigle and P. Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B*, 74:267–286, 2012.

[16] I. Epifanio. Shape descriptors for classification of functional data. *Technometrics*, 50:284–294, 2008.

[17] M.G. Estévez-Pérez and J. A. Vilar. Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics*, 20:495–517, 2013.

[18] J. Fan and S.-K. Lin. Test of significance when data are curves. *Journal of the American Statistical Association*, 93:1007–1021, 1998.

[19] F. Ferraty and P. Vieu. Curves discrimination: A nonparametric functional approach. *Computational Statistics and Data Analysis*, 51:4878–4890, 2003.

[20] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer, New York, 2006.

[21] P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52:1694–1711, 2008.

[22] C. Fraley and A.E. Raftery. Model-Based Clustering,Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association,*, 97:611–631, 2002.

[23] P. Galeano, E. Joseph, and R. E. Lillo. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 2014.

[24] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B*, 68:109–126, 2006.

[25] P. Hall, D. Poskitt, and B. Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9, 2001.

[26] P. Hall and I. Van Keilegom. Two-sample test in functional data analysis starting from discrete data. *Statistica Sinica*, 17:1511–1531, 2007.

[27] W. Härdle. *Applied nonparametric regression*. Cambridge University Press, UK, 1990.

[28] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[29] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

[30] L. Horváth, P. Kokoszka, and R. Reeder. Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B*, 74:533–550, 2013.

[31] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B*, 63:533–550, 2001.

[32] G. S. James. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41:19–43, 1954.

[33] S. Johansen. The welch-james approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67:85–92, 1980.

[34] S. Kim. A practical solution to the multivariate behrens-fisher problem. *Biometrika*, 79:171–176, 1992.

[35] X. Y. Leng and H.-G. Müller. Classification using functional data analysis for temporal gene expression data. *Journal of the Royal Statistical Society: Series B*, 22:68–76, 2006.

[36] B. Li and Q. Yu. Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis*, 52:4790–4800, 2008.

[37] Y. Li, N. Wang, and R.J. Carroll. Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108:1284–1294, 2013.

[38] S. López-Pintado and J. Romo. Depth-based classification for functional data. *In DIMACS Series in Discrete Mathematics and Theoretical Computer Science , American Mathematical Society*, 72:103–120, 2006.

[39] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Academy of Sciences, India*, 12:49–55, 1936.

[40] R. Maronna and R. Zamar. Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, 44(4):307–317, 2002.

[41] B. Martín-Barragán, R. E. Lillo, and J. Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232:146–155, 2014.

[42] A. Mas. Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis*, 98:1231–1261, 2007.

[43] V. Melnykov and I. Melnykov. On k-means algorithm with the use of mahalanobis distances. *Statistics & Probability Letters*, 84:88–95, 2014.

[44] E. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[45] D. G. Nel and C. A. van der Merwe. A solution to the multivariate behrens-fisher problem. *Communications in Statistics-Series A, Theory and Methods*, 15:3719–3735, 1986.

[46] C. Preda and G. Saporta. Pls regression on a stochastic process. *Computational Statistics and Data Analysis*, 48:149–158, 2005a.

[47] J. O. Ramsay, G. Hooker, and S. Graves. *fda: Functional Data Analysis*. package version 2.3.6, URL http://CRAN.R-project.org/package=fda, 2009.

[48] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis: Second Edition*. Springer, New York, 2005.

[49] A. C. Rencher. *Multivariate Statistical Inference and Applications*. Willey, New York, 1998.

[50] A. C. Rencher. *Methods of Multivariate Analysis, Second Edition*. Willey, New York, 2000.

[51] D. Rocke and D. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, 1996.

[52] F. Rossi and Villa N. Support vector machine for functional data classification. *Neurocomputing*, 69:730–742, 2006.

[53] P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[54] P.J. Rousseeuw and B.C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651, 1990.

[55] C. Sguera, P. Galeano, and R. E. Lillo. Spatial depth-based classification for functional data. 2012.

[56] H. Shin. An extension of fisher discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, 99:1191–1216, 2008.

[57] A. J. Smola and R. Kondor. Kernels and regularization on graphs. *In Learning Theory and Kernel Machines, Lectures Notes in Computer Science: Springer-Verlag Berlin*, 2777:144–158, 2003.

[58] X. H. Wang, S. Ray, and B. K. Mallick. Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102:962–973, 2007.

[59] G. Watson. Smooth regression analysis. *Sankhya, Ser. A*, 26:359–372, 1964.

[60] D. Woodruff and D. Rocke. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association*, 89(427):888–896, 1994.

[61] Y. Yao. An approximate degrees of freedom solution to the multivariate behrens-fisher problem. *Biometrika*, 52:139–147, 1965.

[62] C. Q. Zhang, H. Peng, and J. T. Zhang. Two sample tests for functional data. *Comm. in Stat.-Theory and Methods*, 39:559–578, 2010.

[63] J. T. Zhang and J. W. Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35:1052–1079, 2007.

[64] J.T. Zhang, X. Liang, and S. Xiao. On the two-sample behrens-fisher problem for functional data. *Journal of Statistical Theory and Practice*, 4(4), 2011.