View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universidad Carlos III de Madrid e-Archivo

UNIVERSIDAD CARLOS III DE MADRID

working papers

Working Paper 09 - 39
Statistics and Econometrics Series 15
July 2009

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

CLASSIFICATION OF FUNCTIONAL DATA: A WEIGHTED DISTANCE APPROACH

Andrés M. Alonso[1], David Casado[2] and Juan Romo[3]

Abstract

A popular approach for classifying functional data is based on the distances from the function or its derivatives to group representative (usually the mean) functions or their derivatives. In this paper, we propose using a combination of those distances. Simulation studies show that our procedure performs very well, resulting in smaller testing classification errors. Applications to real data show that our procedure performs as well as –and in some cases better than– other classification methods.

Keywords: discriminant analysis, functional data, weighted distances.

[1] Departamento de Estadística. Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: andres.alonso@uc3m.es

[2] Departamento de Estadística. Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: david.casado@uc3m.es

[3] Departamento de Estadística. Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: juan.romo@uc3m.es

1 Introduction

Functional data have great —and growing— importance in Statistics. Most of the classical techniques for the finite- and high-dimensional frameworks have been adapted to cope with the infinite dimensions, but due to the *curse of dimensionality*, new and specific treatments are still required. As with other type of data, statisticians must supervise different steps —registration, missing data, representation, transformation, typicality— and tackle different tasks —modelization, discrimination or clustering, among others. In practice, curves can neither be registered continuously nor at infinite points. Then, techniques dealing with high-dimensional data can sometimes be applied: Hastie et al. (1995), for example, adapt the discriminant analysis to cope with many highly correlated predictors, "such as those obtained by discretizing a function".

Among the approaches specifically designed for functional data classification, the following project the data into a finite-dimensional space of functions and therefore work with the coefficients; this technique is called *filtering*. James and Hastie (2001) model the coefficients with "Gaussian distribution with common covariance matrix for all classes, by analogy with LDA [linear discriminant analysis]"; their classification minimizes the distance to the group mean. The classification method of Hall et al. (2001) maximizes the likelihood, and although they propose a fully nonparametric density estimation, in practice multivariate Gaussian densities are considered, leading to quadratic discriminant analysis. Biau et al. (2003) apply k-nearest neighbour to the coefficients, while Rossi and Villa (2006) apply support vector machines. Berlinet et al. (2008) extend the approach of Biau et al. (2003) to wavelet bases and to more general discrimination rules. The following proposals are designed to make direct use of the continuity of the functional data. Ferraty and Vieu (2003) classify new curves in the group with the highest posterior probability of membership kernel estimate. On the other hand, López-Pintado and Romo (2006) also take into account the continuity feature of the data and propose two classification methods based on the notion of *depth* for curves; in their first proposal new curves are assigned to the group with the closest trimmed mean, while the second method minimizes a weighted average distance to each element in the group. Abraham et al. (2006) extend the moving window rule for functional data classification. Nerini and Ghattas (2007) classify density functions with functional regression trees. Baíllo and Cuevas (2008) provide some theoretical results on the functional k-nearest neighbour classifier, and suggest —as a partial answer— that this method could play the same central role for functional data as Fisher's method for the finite-dimensional case. To use only the most informative parts of the curves, Li and Yu (2008) have proposed a new idea: they use F-statistics to select the place where linear discriminant analysis is applied into small intervals, providing an output that is used as input in a final support vector machines step.

There are several works addressing the unsupervised classification or clustering problem. Abraham et al. (2003) fit the functional data by B-splines and apply k-means on the coefficients. James and Sugar (2003) project the data into a finite-dimensional space and consider a random-effects model for the coefficients; their method is effective when the observations are sparse, irregularly spaced or occur at different time points for each subject. The continuity nature of the data is used, in a more direct form, by the following works. The proposal of Tarpey and Kinateder (2003) classifies using a k-means algorithm over the probability distributions. A hierarchical descending procedure, using heterogeneity indexes based on modal and mean curves, is presented in Dabo-Niang et al. (2006). Impartial trimming is combined with k-means in Cuesta-Albertos and Fraiman (2007).

Functional data can be transformed in several ways. After the registration, spatial or temporal alignments are sometimes necessary; references on this topic are Wang and Gasser (1997), Wang and Gasser (1999) and Ramsay and Silverman (2006). On the other hand, Dabo-Niang et al. (2007) use a distance invariant to small shifts. Examples of centering, normalization and derivative transformations are found in Rossi and Villa (2006). The objective of the transformations is to highlight some features of the data and to allow the information to be used more efficiently. For this kind of data, when they are smooth enough, the most important transformation is the derivation. Since the different derivatives can contribute new information, a possible combination of them —or their information— should be taken into account. Mathematical *Functional Analysis* has been working with such combinations for a long time, mainly through some norms (in norm and Sobolev spaces), and Ramsay and Silverman (2006) find them frequently as a consequence of model adjustments or system properties (for Canadian weather stations data, melanoma data or lower lip movement data).

In order to obtain semimetrics, instead of metrics, Ferraty and Vieu (2006) consider derivatives (one at a time) in the distances. This implies theoretical advantages —throughout the topological structure induced by the semimetric— in the small ball probability function, providing a new way to deal with the curse of dimensionality.

We transform the functional data classification problem into a classical multivariate data classification problem. While the filtering techniques encapsulate the functional information into a set of coefficients, we construct a lineal combination of variables and coefficients. Given the variables, the *linear discriminant analysis* determines the combination. Our proposal is based on the interpretation as variables of the distances between a new curve and the transformed and untransformed functional data. On the one hand, the classification can be improved, and, on the other hand, the coefficients of the combination provide information about the importance of each data transformation. When a nonnegativeness condition is applied to the coefficients, the combination (discriminant function) can be interpreted as the difference of measurements with a weighted distance. This metric automatically becomes a semimetric when the importance of the distance to the untransformed data is null or not significant; but the user can decide, by considering as input only the derivatives, that the methods output necessarily a semimetric.

The paper is organized as follows. In section 2 the classification method is presented and described, from the optimization problem to the classification algorithm. In section 3, our proposal is evaluated with several simulation exercises. Two real data sets are classified in section 4. Finally, in section 5 a summary of conclusions is given.

2 Classification Method

2.1 Motivation

Let $\mathbf{X} = (X_1, \dots, X_p)^t$ be a random vector with mean $\mu_{\mathbf{X}} = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_p])^t$ and covariance matrix $\mathbf{\Sigma_X} = (\sigma_{ij}) = (cov[X_i, X_j]) = \mathbb{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^t]$; when there are $P^{(k)}$, $k = 1, \dots, K$, populations where the vector distribution is $\mathbf{X}^{(k)} = (X_1^{(k)}, \dots, X_p^{(k)})^t$, with parameters $\mu_{\mathbf{X}}^{(k)}$ and $\mathbf{\Sigma}_{\mathbf{X}}^{(k)}$, respectively, capturing the differences between the groups from the distribution of $\mathbf{X}$ is a subject of great interest.

On the other hand, it is frequently convenient or necessary to summarize the information of a vector in a shorter one; that is, to consider $\mathbf{Y} = (Y_1, \dots, Y_q)^t$, with $q < p$, instead of $\mathbf{X} = (X_1, \dots, X_p)^t$.

The previous two tasks can be done simultaneously via the following multiple transformation, where the coefficients can be interpreted as weights (in the sense explained in section 2.2.3):

$$Y_j^{(k)} = a_{j1} X_1^{(k)} + \dots + a_{jp} X_p^{(k)}, \qquad j = 1, \dots, q, \tag{1}$$

or, in matrix notation,

$$\mathbf{Y}^{(k)} = \mathbf{A}\mathbf{X}^{(k)}, \tag{2}$$

where $\mathbf{A}$ is the $q \times p$ matrix of the coefficients. Notice that $\mathbf{A}$ is independent of k, that is, independent of the population. The superscript (k) has been maintained in the notation to highlight that the new vector $\mathbf{Y}$ also has a different distribution in each population, and that the election of $\mathbf{A}$ must preserve or increase this difference so that $\mathbf{Y}$ is suitable for discrimination. The covariance matrix of $\mathbf{Y}^{(k)}$ is $\mathbf{\Sigma}_{\mathbf{Y}}^{(k)} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{X}}^{(k)}\mathbf{A}^t$.

Let us consider, for each population k, the sample

$$(\mathbf{x}_1^{(k)}, \cdots, \mathbf{x}_{n_k}^{(k)}) = \begin{pmatrix} x_{11}^{(k)} & \cdots & x_{n_k 1}^{(k)} \\ \vdots & \ddots & \vdots \\ x_{1p}^{(k)} & \cdots & x_{n_k p}^{(k)} \end{pmatrix}, \quad k = 1, \dots, K, \tag{3}$$

where $\mathbf{x}_j^{(k)}$, the j-th column of the matrix, contains the j-th element of the sample, and $n = \sum_{k=1}^{K} n_k$. In the following subsections some known theory of this multivariate framework is given in order to motivate Fisher's method criterion.

2.1.1 Parameter Estimation

The parameters of the distributions can be estimated from samples as follows. The quantity $\overline{x}_i^{(k)} = n_k^{-1} \sum_{j=1}^{n_k} x_{ij}^{(k)}$ estimates $\mathbb{E}(X_i^{(k)})$, while the quantity $\overline{\mathbf{x}}^{(k)} = n_k^{-1} \sum_{j=1}^{n_k} \mathbf{x}_j^{(k)}$ estimates $\mu_{\mathbf{X}}^{(k)}$.

The matrixes $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(k)} = n_k^{-1}\sum_{j=1}^{n_k}(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}}^{(k)})(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}}^{(k)})^t$ and $\mathbf{S}_{\mathbf{x}}^{(k)} = \frac{n_k}{n_k-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(k)}$ estimate $\mathbf{\Sigma}_{\mathbf{X}}^{(k)}$ with and without bias, respectively. The covariance matrix $\mathbf{\Sigma}_{\mathbf{Y}}^{(k)}$ is estimated, with and without bias, respectively, by $\hat{\mathbf{\Sigma}}_{\mathbf{y}}^{(k)} = \mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(k)}\mathbf{A}^t$ and $\mathbf{S}_{\mathbf{y}}^{(k)} = \mathbf{A}\mathbf{S}_{\mathbf{x}}^{(k)}\mathbf{A}^t$.

When $\mathbf{\Sigma}_{\mathbf{X}}^{(k)} = \mathbf{\Sigma}_{\mathbf{X}}$ for all k, the matrix $\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \sum_{k=1}^{K}\frac{n_k}{n}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(k)}$ estimates $\mathbf{\Sigma}_{\mathbf{X}}$ with bias, while an unbiased estimator is $\mathbf{S}_{\mathbf{x}} = \frac{n}{n-K}\hat{\mathbf{\Sigma}}_{\mathbf{x}}$.

2.1.2 Variability Information

Information about the within- and between-group variabilities are provided, respectively, by the *within-class scatter matrix*

$$\mathbf{W} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}}^{(k)})(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}}^{(k)})^t, \tag{4}$$

and the *between-class scatter matrix*

$$\mathbf{B} = \sum_{k=1}^{K} n_k(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})^t, \tag{5}$$

where $\overline{\mathbf{x}} = n^{-1}\sum_{k=1}^{K} n_k\overline{\mathbf{x}}^{(k)}$ is the global mean. The *total scatter matrix*,

$$\mathbf{T} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}})(\mathbf{x}_j^{(k)} - \overline{\mathbf{x}})^t, \tag{6}$$

expresses the total variability and is the sum of the previous quantities, $\mathbf{W} + \mathbf{B} = \mathbf{T}$.

These three matrixes are symmetric by definition. In addition, $\mathbf{W}$ is full rank and positive definite, while $\mathbf{B}$ is positive semidefinite.

Remark 1: The discriminant analysis is a supervised classification technique where the membership information is exploited through these variability matrixes.

Remark 2: An important observation is that from definition (4) it holds that

$$\mathbf{W} = \sum_{k=1}^{K} n_k\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(k)} = (n-K)\mathbf{S}_{\mathbf{x}}. \tag{7}$$

This implies that both matrixes, $\mathbf{W}$ and $\mathbf{S}_{\mathbf{x}}$, could be used in the statements of this document. A positive constant factor does not change the optimization problems that will be considered. Nevertheless, we shall use $\mathbf{W}$ since it maintains its meaning as variability matrix, while the matrix $\mathbf{S}_{\mathbf{x}}$ makes sense as an estimator only under the fulfilment of the equal group variability assumption (homoscedasticity).

2.1.3 Splitting Criterion

Given a unique sample with elements of both populations, minimizing a functional of $\mathbf{W}$ or maximizing a functional of $\mathbf{B}$ is a reasonable criterion for splitting the sample from the information provided by the vector $\mathbf{x}$. Many techniques are based on this idea.

2.1.4 Case $q = 1$: One Function

Some methods in the literature choose —with different criteria— the linear combinations $y_j^{(k)}$ one at a time, so the case $q = 1$ is specially considered therefore:

$$y^{(k)} = a_1 x_1^{(k)} + \ldots + a_p x_p^{(k)} = \mathbf{a}^t \mathbf{x}^{(k)}, \tag{8}$$

with $\mathbf{a} = (a_1, \ldots, a_p)^t$. For this new compound variable $\overline{y}^{(k)} = \mathbf{a}^t \overline{\mathbf{x}}^{(k)}$ and $(s_y^{(k)})^2 = \mathbf{a}^t \mathbf{S}_{\mathbf{x}}^{(k)} \mathbf{a}$, where $\mathbf{S}_{\mathbf{x}}^{(k)}$ is the within-group sample covariance matrix. For classifying purposes, the variable $y^{(k)}$ must discriminate as much as possible. Following the idea of the above-mentioned splitting criterion, the interest is in finding $\mathbf{a}$ so as to minimize the within-group dispersion,

$$\begin{aligned}
W_y &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_i^{(k)} - \overline{y}^{(k)})(y_i^{(k)} - \overline{y}^{(k)})^t \\
&= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{a}^t \mathbf{x}_i^{(k)} - \mathbf{a}^t \overline{\mathbf{x}}^{(k)})(\mathbf{a}^t \mathbf{x}_i^{(k)} - \mathbf{a}^t \overline{\mathbf{x}}^{(k)})^t \\
&= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{a}^t (\mathbf{x}_i^{(k)} - \overline{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \overline{\mathbf{x}}^{(k)})^t \mathbf{a} \\
&= \mathbf{a}^t \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \overline{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \overline{\mathbf{x}}^{(k)})^t \right] \mathbf{a} = \mathbf{a}^t \mathbf{W} \mathbf{a}, \qquad (9)
\end{aligned}$$

or to maximize the between-group dispersion,

$$\begin{aligned}
B_y &= \sum_{k=1}^{K} n_k (\overline{y}^{(k)} - \overline{y})(\overline{y}^{(k)} - \overline{y})^t \\
&= \sum_{k=1}^{K} n_k (\mathbf{a}^t \overline{\mathbf{x}}^{(k)} - \mathbf{a}^t \overline{\mathbf{x}})(\mathbf{a}^t \overline{\mathbf{x}}^{(k)} - \mathbf{a}^t \overline{\mathbf{x}})^t \\
&= \sum_{k=1}^{K} n_k \mathbf{a}^t (\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})^t \mathbf{a} \\
&= \mathbf{a}^t \left[\sum_{k=1}^{K} n_k (\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})^t \right] \mathbf{a} = \mathbf{a}^t \mathbf{B} \mathbf{a}. \qquad (10)
\end{aligned}$$

Then, the election of $\mathbf{a}$ can be formulated as an *optimal weighting problem.*

2.2 The Optimization Problem

Fisher's proposal is based on a tradeoff criterion between the two previous ones, as it maximizes the —sometimes termed— *generalized Rayleigh quotient*: $\lambda = B_y / W_y$. To add each consecutive compound function, this method also maximizes this quantity but with the imposition of incorrelation with the previous combination. Another interesting interpretation arises when the generalized Rayleigh quotient is written as

$$\lambda(\mathbf{a}) = \frac{B_y}{W_y} = \frac{\mathbf{a}^t\mathbf{B}\mathbf{a}}{\mathbf{a}^t\mathbf{W}\mathbf{a}} = \frac{\mathbf{a}^t(\mathbf{B}+\mathbf{W}-\mathbf{W})\mathbf{a}}{\mathbf{a}^t\mathbf{W}\mathbf{a}} = \frac{\mathbf{a}^t\mathbf{T}\mathbf{a}}{\mathbf{a}^t\mathbf{W}\mathbf{a}} - 1. \tag{11}$$

This decomposition shows that maximizing $\lambda(\mathbf{a})$ can be interpreted as maximizing the total variability while minimizing the within-class variability.

Since our procedure is based on the use of Fisher's discriminant analysis, with one ($q = 1$) *discriminant function* y and several ($p \geq 1$) *discriminant variables* $x_1, \ldots, x_p$, our optimization problem consists of finding $\mathbf{a}$ such that:

$$\mathbf{a} = argmax \left\{\lambda(\mathbf{a})\right\} = argmax \left\{\frac{B_y}{W_y}\right\} = argmax \left\{\frac{\mathbf{a}^t\mathbf{B}\mathbf{a}}{\mathbf{a}^t\mathbf{W}\mathbf{a}}\right\}. \tag{12}$$

This is a *nonlinear (quadratic) optimization problem.* As $\lambda(c\mathbf{a}) = \lambda(\mathbf{a}) \;\; \forall c \in \mathbb{R}, \;\; c \neq 0$, the solution —when it exists— will not be a unique vector but an infinite family of them, denoted by $E^*_\lambda = \{c\mathbf{a} \mid c \in \mathbb{R}, \;\; c \neq 0\}$.

The analytical resolution of the problem is obtained by solving:

$$\frac{\partial\lambda}{\partial\mathbf{a}} = \frac{2[\mathbf{B}\mathbf{a}(\mathbf{a}^t\mathbf{W}\mathbf{a}) - (\mathbf{a}^t\mathbf{B}\mathbf{a})\mathbf{W}\mathbf{a}]}{(\mathbf{a}^t\mathbf{W}\mathbf{a})^2} = \frac{2[\mathbf{B}\mathbf{a} - \lambda\mathbf{W}\mathbf{a})]}{\mathbf{a}^t\mathbf{W}\mathbf{a}} = \mathbf{0}, \tag{13}$$

so

$$\mathbf{B}\mathbf{a} - \lambda\mathbf{W}\mathbf{a} = (\mathbf{B} - \lambda\mathbf{W})\mathbf{a} = \mathbf{0} \tag{14}$$

is the eigenequation of the problem; for a nonnull solution to exist, it is necessary that $|\mathbf{B}-\lambda\mathbf{W}| = 0$. As $\mathbf{W}$ be invertible (not singular),

$$\mathbf{W}^{-1}(\mathbf{B} - \lambda\mathbf{W})\mathbf{a} = (\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}. \tag{15}$$

The interest is in the largest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$. If $\mathbf{a}$ is a nonnull eigenvector of λ, so is any element of E^*_λ; that is, the set of eigenvectors is solution of (15). Let us denote the solution of this optimization problem by the pair $(E_{\lambda_F}, \lambda_F)$.

To avoid the arbitrary scale factor and obtain a unique solution, usually a constraint is added:

$$\mathbf{a} = argmax \left\{\mathbf{a}^t\mathbf{B}\mathbf{a}\right\} \quad \text{subject to} \quad \mathbf{a}^t\mathbf{W}\mathbf{a} = 1 \tag{16}$$

so that the solution is $(\mathbf{a}_F, \lambda_F)$, with $\mathbf{a}^t_F\mathbf{W}\mathbf{a}_F = 1$ and $\lambda_F = \mathbf{a}^t_F\mathbf{B}\mathbf{a}_F$. The computes leading to the explicit expression of $\mathbf{a}_F$ are given in section 2.3.

As a final general, theoretical comment, when the distributions of $\mathbf{X}^{(k)}$ are normal, Fisher's approach is optimal in the sense of minimizing the misclassification probability.

2.2.1 An Additional Constraint

In order to base the classification on a semimetric or on a metric, one version of our proposal adds another constraint —in fact, several nonnegativity constraints— to the optimization problem:

$$\mathbf{a} = \mathit{argmax}\left\{\mathbf{a}^t\mathbf{B}\mathbf{a}\right\} \quad \text{subject to} \quad \begin{cases} \mathbf{a}^t\mathbf{W}\mathbf{a} = 1 \\ \mathbf{a} \geq \mathbf{0} \end{cases}, \tag{17}$$

where $\mathbf{B}$ is the between-class scatter matrix, $\mathbf{W}$ is the within-class scatter matrix, $\mathbf{a} = (a_1, \dots, a_p)^t$, and $\mathbf{a} \geq \mathbf{0}$ means $a_i \geq 0$, for $i = 1, \dots, p$. This is a nonlinear (quadratic) programming problem with an *equality constraint* and *nonnegativity constraints.* The solution of this new optimization problem can be represented by the pair $(\mathbf{a}_p, \lambda_p)$, with $\mathbf{a}_p^t\mathbf{W}\mathbf{a}_p = 1$, $\mathbf{a}_p \geq \mathbf{0}$ and $\lambda_p = \mathbf{a}_p^t\mathbf{B}\mathbf{a}_p$. Although the optimization problem can be solved computationally, section 2.3 contains some theory on obtaining the explicit expression of $\mathbf{a}_p$.

Geometrically, the set $E_\lambda = E_\lambda^* \cup \{\mathbf{a} = \mathbf{0}\}$ is a one-dimensional vectorial subspace of $\mathbb{R}^p$. When E_λ intersects the nonnegative orthant $\{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a} \geq \mathbf{0}\}$ outside the origin, this last optimization problem will provide the same solution as those without the nonnegativity constraints.

2.2.2 Case $K = 2$: Two Populations

We have considered the classification into two populations. It is well-known that this case can be written as an equivalent linear regression problem; nevertheless, we have not used this interpretation.

Since

$$(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}) = \frac{n_2}{n}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}) \tag{18}$$

and

$$(\overline{\mathbf{x}}^{(2)} - \overline{\mathbf{x}}) = \frac{n_1}{n}(\overline{\mathbf{x}}^{(2)} - \overline{\mathbf{x}}^{(1)}) = -\frac{n_1}{n}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}), \tag{19}$$

the term $\mathbf{a}^t\mathbf{B}\mathbf{a}$ can be expressed as

$$\begin{aligned} \mathbf{a}^t\mathbf{B}\mathbf{a} &= \mathbf{a}^t\left(\sum_{k=1}^{K} n_k(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}})^t\right)\mathbf{a} \\ &= \mathbf{a}^t\frac{n_1n_2}{n}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{a} \\ &= \frac{n_1n_2}{n}[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})][\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^t \\ &= \frac{n_1n_2}{n}[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2. \end{aligned} \tag{20}$$

Furthermore, when the population covariance matrixes are supposed to be equal, the sample information can be combined to estimate the common covariance matrix, and the previous optimization problems are equivalent, respectively, to the following ones:

$$\mathbf{a} = argmax \left\{ \frac{[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t\mathbf{W}\mathbf{a}} \right\} \tag{21}$$

or

$$\mathbf{a} = argmax \left\{ [\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2 \right\} \qquad \text{subject to} \qquad \mathbf{a}^t\mathbf{W}\mathbf{a} = 1, \tag{22}$$

and, finally, with our additional restriction,

$$\mathbf{a} = argmax \left\{ [\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2 \right\} \qquad \text{subject to} \qquad \left\{ \begin{array}{l} \mathbf{a}^t\mathbf{W}\mathbf{a} = 1 \\ \mathbf{a} \geq \mathbf{0} \end{array} \right. . \tag{23}$$

We have implemented the two last versions. Notice that with this formulation the numerator highlights the objective of the optimization problem — maximizing the difference between the means under control of the variability.

2.2.3 Interpretation of the Coefficients

Usually the variables of the vector $\mathbf{x}$ have been measured using different scales: localization, variability or even units of measure. Then, the mathematical solution of the optimization problem provides values a_i not taking into account this fact. The function

$$y = a_1x_1 + \ldots + a_px_p = \mathbf{a}^t\mathbf{x}, \tag{24}$$

with $\mathbf{a} = (a_1, \ldots, a_p)^t$, can, however, be used for classifying.

A possible transformation that can be applied to the previous values is: a translation so that the axes origin coincide with the global centroid of the samples and an homothecy so that the coordinates be referred to the standard deviation of each axis. Then,

$$\tilde{y} = b_0 + b_1x_1 + \ldots + b_px_p = b_0 + \mathbf{b}^t\mathbf{x}, \tag{25}$$

with $\mathbf{b} = (b_1, \ldots, b_p)^t$, where b_i can be interpreted as regression coefficients. When these values are computed from crude data, they are termed *nonstandardized coefficients*, that represent the (absolute) contribution of the variables to the function but are not comparable among them. Anyway, $\tilde{y}$ is also used for classifying. On the other hand, the typification of variables solves at the same time the aforementioned scale problems —localization, variability and units—, so if $\mathbf{b}$ is computed from typified —not crude— variables, this vector contents the *standardized coefficients*, that represent the relative contribution of the variables to the function and are comparable among them. Nevertheless, now the function (25) cannot be used for classifying, as the important information has been lost (in this case $b_0 = 0$, for example).

Our proposal has been described in terms of the function (24) suggested by the optimization problem; nevertheless, the interpretation of the coefficients —through the figures— has been based on the function

$$y = \mathbf{a}^t\mathbf{x} = \mathbf{a}^t\mathbf{D}\mathbf{D}^{-1}\mathbf{x} = \mathbf{a}^t\mathbf{D}\tilde{\mathbf{x}}, \tag{26}$$

where $\tilde{\mathbf{x}} = \mathbf{D}^{-1}\mathbf{x}$, with $\mathbf{D}$ being the diagonal matrix with elements $\sigma_1, \ldots, \sigma_p$, where σ_i is the standard deviation of the variable x_i. After applying this *univariate standardization*, the new variables have variance equal to one:

$$\tilde{\mathbf{x}} = \mathbf{D}^{-1}\mathbf{x} = (\sigma_1^{-1}x_1, \ldots, \sigma_p^{-1}x_p) \quad \Rightarrow \quad Var(\tilde{x_i}) = Var(\sigma_i^{-1}x_i) = 1. \tag{27}$$

Note that the previous transformation does not change the mean of each variable. Thus, for the interpretation we have considered the coefficients defined by $\mathbf{a}^t\mathbf{D}$, that is, the quantities

$$\mathbf{a}^t\mathbf{D} = (a_1\sigma_1, \ldots, a_p\sigma_p). \tag{28}$$

2.3 The Discriminant Function

For two populations, the resolution of the optimization problem, with and without the classical constraint (when $\mathbf{a}^t\mathbf{W}\mathbf{a} = 1$ or $\beta = 0$, respectively), is given at the same time by

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \mathbf{a}}\left(\frac{[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t\mathbf{W}\mathbf{a}} - \beta(\mathbf{a}^t\mathbf{W}\mathbf{a} - 1)\right) \\
&= \frac{\partial}{\partial \mathbf{a}}\left(\frac{[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t\mathbf{W}\mathbf{a}}\right) - \beta 2\mathbf{W}\mathbf{a} \\
&= \frac{2\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})\mathbf{a}^t\mathbf{W}\mathbf{a} - [\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2 2\mathbf{W}\mathbf{a}}{(\mathbf{a}^t\mathbf{W}\mathbf{a})^2} \\
&\quad -\beta 2\mathbf{W}\mathbf{a}
\end{aligned} \tag{29}$$

so

$$\frac{\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})\mathbf{a}^t\mathbf{W}\mathbf{a}}{[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2 + \beta(\mathbf{a}^t\mathbf{W}\mathbf{a})^2}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}) = \mathbf{W}\mathbf{a} \tag{30}$$

and, when $\mathbf{W}$ is invertible (not singular),

$$\mathbf{a} = \frac{\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})\mathbf{a}^t\mathbf{W}\mathbf{a}}{[\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})]^2 + \beta(\mathbf{a}^t\mathbf{W}\mathbf{a})^2}\mathbf{W}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}). \tag{31}$$

As $\mathbf{a}^t(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$ and $\mathbf{a}^t\mathbf{W}\mathbf{a}$ are numbers, it does not matter whether the constraint $\mathbf{a}^t\mathbf{W}\mathbf{a} = 1$ is imposed or not, the solution for the classical linear discriminant analysis is that y is proportional to $(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}$, and, without loss of generality:

$$y = \mathbf{a}_F^t\mathbf{x} = (\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}\mathbf{x}. \tag{32}$$

Since $y \in \mathbb{R}$, it is sometimes written as $y = y^t = \mathbf{x}^t\mathbf{a}_F = \mathbf{x}^t\mathbf{W}^{-1}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})$ in the literature.

The expression of the discriminant function with our additional constraint, $y = \mathbf{a}_p^t\mathbf{x}$, is more difficult to obtain. We present explicit expressions for some specific easy cases (notice that in this work we consider the cases $p = 1$, 2 or 3). Although we are interested in the $K = 2$ case, some of the following calculations are done with the same difficulty for the general K-populations case: that is, for the general problem (16) instead of the particular one (22). As was mentioned, when the vectorial subspace E_λ of $\mathbb{R}^p$ intersects the nonnegative orthant $\{\mathbf{a} \in \mathbb{R}^p \ / \ \mathbf{a} \geq \mathbf{0}\}$ outside the origin, that is, when all the components of $\mathbf{a}_F$ have the same sign, the new discriminant function will be

$$y = \mathbf{a}_p^t \mathbf{x} = \alpha \mathbf{a}_F^t \mathbf{x} = \alpha(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}\mathbf{x}, \tag{33}$$

with $\alpha = +1$ or $\alpha = -1$ so that the condition $\alpha \mathbf{a}_F \geq \mathbf{0}$ holds.

In general, when all the components of $\mathbf{a}_F$ do not have the same sign, formal calculations are necessary. The *objective function* —of the optimization problem— and the constraints are combined in the *Lagrangian*, and the nonnegativeness is taken into account through the *Karush--Kuhn-Tucker conditions*:

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{a}} = \mathbf{0} \\[2ex] \dfrac{\partial L}{\partial \beta} = \mathbf{0} \\[2ex] a_i \geq 0, \ \mu_i \geq 0 \ \text{ and } \ \mu_i a_i = 0 \end{cases} , \tag{34}$$

where the Lagrangian is

$$L(\mathbf{a}, \beta, \mu) = \mathbf{a}^t \mathbf{B} \mathbf{a} - \beta(\mathbf{a}^t \mathbf{W} \mathbf{a} - 1) + \mathbf{a}^t \mu \tag{35}$$

and $\mu = (\mu_i, \ldots, \mu_p)^t$ and β are the multipliers. It holds that

$$\frac{\partial L}{\partial \mathbf{a}} = 2\mathbf{B}\mathbf{a} - \beta 2\mathbf{W}\mathbf{a} + \mu. \tag{36}$$

The conditions (34) become

$$\begin{cases} 2(\mathbf{B} - \beta \mathbf{W})\mathbf{a} = -\mu \\[2ex] \mathbf{a}^t \mathbf{W} \mathbf{a} = 1 \\[2ex] a_i \geq 0, \ \mu_i \geq 0 \ \text{ and } \ \mu_i a_i = 0, \end{cases} \tag{37}$$

that are a system with $2p + 1$ conditions and variables. Giving explicit solution of this system is only possible in some simple cases.

2.3.1 Case $p = 1$: One Variable

This case, with only one discriminant variable, is trivial since

$$\lambda(a) = \frac{aBa}{aWa} = \frac{B}{W} = constant. \tag{38}$$

2.3.2 Case $p = 2$: Two Variables

First of all, when two populations are considered, let us denote

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

where by definition $w_{12} = w_{21}$ and $b_{12} = b_{21}$.

For two discriminant variables, three nonnull subcases arise from the Karush-Kuhn-Tucker conditions.

(A) Case $a_1 > 0,\ \mu_1 = 0$ and $a_2 = 0$. In this case,

(a1) By hypothesis, $\mu_1 = 0$ and $a_2 = 0$.

(a2) From $\mathbf{a}^t\mathbf{W}\mathbf{a} = 1$ the value $a_1 = |\sqrt{w_{11}^{-1}}|$ is obtained.

(a3) Finally, $2(\mathbf{B}-\beta\mathbf{W})\mathbf{a} = -\mu$ implies that $\beta = w_{11}^{-1}b_{11}$ and $\mu_2 = -2(b_{21}-w_{11}^{-1}b_{11}w_{21})|\sqrt{w_{11}^{-1}}|$.

The discriminant function, if $\mathbf{B} - \beta\mathbf{W}$ is negative semidefinite, would be

$$y_A = \mathbf{a}_p^t\mathbf{x} = |\sqrt{w_{11}^{-1}}|x_1. \tag{39}$$

(B) Case $a_1 = 0,\ a_2 > 0$ and $\mu_2 = 0$. In this case,

(b1) By hypothesis, $a_1 = 0$ and $\mu_2 = 0$.

(b2) Now, $\mathbf{a}^t\mathbf{W}\mathbf{a} = 1$ implies the value $a_2 = |\sqrt{w_{22}^{-1}}|$.

(b3) From $2(\mathbf{B} - \beta\mathbf{W})\mathbf{a} = -\mu$ the values $\beta = w_{22}^{-1}b_{22}$ and $\mu_1 = -2(b_{12} - w_{22}^{-1}b_{22}w_{12})|\sqrt{w_{22}^{-1}}|$ are obtained.

The discriminant function, if $\mathbf{B} - \beta\mathbf{W}$ is negative semidefinite, would be

$$y_B = \mathbf{a}_p^t\mathbf{x} = |\sqrt{w_{22}^{-1}}|x_2. \tag{40}$$

(C) Case $a_1 > 0,\ \mu_1 = 0,\ a_2 > 0$ and $\mu_2 = 0$. In this case,

(c1) By hypothesis $\mu = \mathbf{0}$, the nonnegativity constraint disappears from the Lagrangian and the objective function is again $L(\mathbf{a}) = \lambda(\mathbf{a})$.

(c2) As $(\mathbf{B} - \beta\mathbf{W})\mathbf{a} = \mathbf{0}$, it is necessary that $|\mathbf{B} - \beta\mathbf{W}| = 0$; this condition implies, since $\mathbf{W}$ is not singular, that

$$\beta = \frac{-b \pm \sqrt{b^2 - 4|\mathbf{W}||\mathbf{B}|}}{2|\mathbf{W}|}, \tag{41}$$

with $b = w_{12}b_{21} + w_{21}b_{12} - w_{11}b_{22} - w_{22}b_{11}$. This means that $(\mathbf{W}^{-1}\mathbf{B} - \beta\mathbf{I})\mathbf{a} = \mathbf{0}$, and we are again interested in an eigenvector of an eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. Nevertheless, now the criterion is not selecting the largest eigenvalue, but selecting the largest one with eigenvectors verifying the nonnegativity constraint (or nonpositiveness, since the scale factor is never a problem).

(c3) Given β, also from $(\mathbf{B}-\beta\mathbf{W})\mathbf{a}=\mathbf{0}$ there will be nontrivial solution if $a_2=\gamma a_1$, or, equivalently, $a_1=\gamma^{-1}a_2$ with

$$\gamma=-\frac{b_{11}-\beta w_{11}}{b_{12}-\beta w_{12}}, \quad \text{or, equivalently,} \quad \gamma=-\frac{b_{21}-\beta w_{21}}{b_{22}-\beta w_{22}}, \tag{42}$$

as $|\mathbf{B}-\beta\mathbf{W}|=0$.

(c4) Finally, the condition $\mathbf{a}^t\mathbf{W}\mathbf{a}=1$ implies that

$$a_1=|\sqrt{[w_{11}+\gamma(w_{12}+w_{21})+\gamma^2 w_{22}]^{-1}}|, \tag{43}$$

or, respectively,

$$a_2=|\sqrt{[\gamma^{-2}w_{11}+\gamma^{-1}(w_{12}+w_{21})+w_{22}]^{-1}}|, \tag{44}$$

so the discriminant function, if $\mathbf{B}-\beta\mathbf{W}$ is negative semidefinite, would be

$$y_C=\mathbf{a}_p^t\mathbf{x}=a_1x_1+\gamma a_1x_2, \tag{45}$$

or, respectively,

$$y_C=\mathbf{a}_p^t\mathbf{x}=\gamma^{-1}a_2x_1+a_2x_2, \tag{46}$$

with γ and β as given above.

Remark 3: In the last computes it has been implicitly supposed that $\gamma\neq 0$ and $\gamma\neq\infty$. Nevertheless, it is noteworthy that when $\gamma\to 0$ or $\gamma\to\infty$, the discriminant functions of the cases (A) and (B) arise, respectively, as limit cases of (C). This is how the parameter γ acquires an important role, since it provides information about each variable importance for classifying purposes, that is, about each variable discriminant power. As $a_1\longrightarrow|\sqrt{w_{11}^{-1}}|$ when $\gamma\to 0$ and $a_2\to|\sqrt{w_{22}^{-1}}|$ when $\gamma\to\infty$, respectively, then

$$\underset{\gamma\to 0}{y_C\longrightarrow y_A} \quad \text{and} \quad \underset{\gamma\to\infty}{y_C\longrightarrow y_B}. \tag{47}$$

Remark 4: This simple case, $p=2$ (two variables), can be used to understand better the meaning of the within-class scatter matrix. By definition,

$$\mathbf{W}=\sum_{k=1}^{K}n_k\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{(k)}=\sum_{k=1}^{K}n_k\left(\hat{\sigma}_{ij}^{(k)}\right)=\left(\sum_{k=1}^{K}n_k\hat{\sigma}_{ij}^{(k)}\right), \tag{48}$$

where $\hat{\sigma}_{ij}^{(k)}=n_k^{-1}\sum_{h=1}^{n_k}(x_{hi}^{(k)}-\overline{x}_i^{(k)})(x_{hj}^{(k)}-\overline{x}_j^{(k)})$. Then, for K populations,

$$\mathbf{W}=(w_{ij})=\left(\sum_{k=1}^{K}\sum_{h=1}^{n_k}(x_{hi}^{(k)}-\overline{x}_i^{(k)})(x_{hj}^{(k)}-\overline{x}_j^{(k)})\right), \tag{49}$$

and, for two populations and two variables,

$$\mathbf{W}=\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22}\end{pmatrix}=\begin{pmatrix} n_1\hat{\sigma}_{11}^{(1)}+n_2\hat{\sigma}_{11}^{(2)} & n_1\hat{\sigma}_{12}^{(1)}+n_2\hat{\sigma}_{12}^{(2)} \\ n_1\hat{\sigma}_{21}^{(1)}+n_2\hat{\sigma}_{21}^{(2)} & n_1\hat{\sigma}_{22}^{(1)}+n_2\hat{\sigma}_{22}^{(2)}\end{pmatrix}. \tag{50}$$

Remark 5: From the Karush-Kuhn-Tucker conditions of the cases $p = 3$ or $p = 4$, several subcases would arise after some work, providing explicit expressions for $\mathbf{a}_p$ under some conditions on the samples. Nevertheless, since it has been proved that there are no formula for the solution of a five-degree general polinomial equation, for the cases $p \geq 5$ it seems impossible to find —in this way— the explicit expressions for $\mathbf{a}_p$.

2.4 The Classification

Geometrically, the classical linear discriminant analysis projects the data into one-dimensional vectorial subspaces. For the first direction, this operation is analytically done by the $y = \mathbf{a}_F^t\mathbf{x}$ operation; the multivariate vector $\mathbf{x}$ is projected by the $(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}$ premultiplication. The method classifies a new element in the population k as follows:

$$\begin{cases} k = 1 & \text{if} \quad y > \frac{1}{2}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)}) \\ \\ k = 2 & \text{otherwise} \end{cases} , \tag{51}$$

where the value $\frac{1}{2}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)})$, the *cutoff point*, is the projection of the midpoint between the two population sample averages, $\frac{1}{2}(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)})$, into the same subspace.

The same ideas apply to the solution obtained with the nonnegativity constraint. Geometrically, the condition $\mathbf{a} \geq 0$ restricts the possible directions into which the data should be projected. We also determine the cutoff point by projecting $\frac{1}{2}(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)})$ with $y = \mathbf{a}_p^t\mathbf{x}$, that is, via the $\mathbf{a}_p^t\cdot$ premultiplication. The method classifies a new element in the population k as follows:

$$\begin{cases} k = 1 & \text{if} \quad y > \frac{1}{2}\mathbf{a}_p^t(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)}) \\ \\ k = 2 & \text{otherwise} \end{cases} , \tag{52}$$

where the value $\frac{1}{2}\mathbf{a}_p^t(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)})$ is the *adjusted cutoff point.* Notice that for the particular case (33) the classification is just the same as that of the classical discriminant analysis.

Thus, for both $y = \mathbf{a}_F^t\mathbf{x}$ and $y = \mathbf{a}_p^t\mathbf{x}$ the classification of a multivariate point is done by the simple comparison of its projection with the projection of the semisum of the group means. The calculations with simulated and real data show that the classification provided by the two discriminant functions is similar, while the nonnegativity restriction adds some theoretical advantages.

Remark 6: The use of data and the previous optimization problem provide a value for $\mathbf{a}$. Then, if there is interest in the stochastic character of the vectors $\mathbf{X}$ and Y (see the motivation at the beginning of this section), the following discriminant function

$$Y = Y(\mathbf{X}) = \mathbf{a}_F^t\mathbf{X} = (\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}\mathbf{X} \tag{53}$$

and classification rule

$$\begin{cases} k = 1 & \text{if} \quad Y > \frac{1}{2}(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)})^t\mathbf{W}^{-1}(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)}) \\ k = 2 & \text{otherwise} \end{cases} \tag{54}$$

can be considered instead of (24) and (51), where $\mathbf{X}$ and Y are random variables again. This can be interpreted as if some population information were unknown and the use of samples would have allowed inferring it.

2.5 Our Discriminant Variables

In order to facilitate understanding of the classification criterion, so far we have used generic discriminant variables $x_1, \ldots, x_p$. Now we define the specific variables and explain how to construct them from the functional data.

If f and g are functions in $C^{p-1}[0,1]$, that is, the set of functions defined in $[0,1]$, being $(p-1)$-times differentiable and with continuous $(p-1)$-derivative; then the quantities $d(f^{i)}, g^{i)})$, for $i = 0, 1, \ldots, p-1$, are numeric when $d(\cdot,\cdot)$ is a distance and the superscript $i)$ denotes the i-th derivative ($i = 0$ represents no differentiation).

Assuming that there are two populations, $K = 2$, with models $f(t)$ and $g(t)$, and let $f_1, \ldots, f_{n_f}$ and $g_1, \ldots, g_{n_g}$ be samples of these populations, respectively; in this situation, for a function $h(t)$ we define the variables

$$x_i = d(h^{i-1)}, \overline{f}^{i-1)}) - d(h^{i-1)}, \overline{g}^{i-1)}), \tag{55}$$

for $i = 1, 2, \ldots, p$, where $\overline{f}^{i-1)} = \frac{1}{n_f}\sum_{j=1}^{n_f} f_j^{i-1)} = (\frac{1}{n_f}\sum_{j=1}^{n_f} f_j)^{i-1)}$ and $\overline{g}^{i-1)} = \frac{1}{n_g}\sum_{j=1}^{n_g} g_j^{i-1)} = (\frac{1}{n_g}\sum_{j=1}^{n_g} g_j)^{i-1)}$. That is, x_i is the difference between the distances from $h^{i-1)}$ to the $(i-1)$-th derivative of the population means.

With these definitions, the discriminant analysis will provide information about the usefulness of each derivative for classification purposes.

2.5.1 Standardization and Coefficients

At this point, it is advisable to study the relation between the variables just defined and the interpretation of the coefficients provided by the optimization problem (see section 2.2.3).

Supposing that a variable t and a function $f(t)$ are not dimensionless (scalars without units of measure), nor is df/dt. Besides, the derivative has a different dimension than its original function, as the term df has the same units than f while the term dt has not them. As a consequence, all the variables defined in (55) are dimensionless only when so are t and f.

Anyway, for classification and descriptive purposes the transformation and the standardization of the data must be applied, respectively, as explained in section 2.2.3. In our methodology this could be done —tried doing— over the functions (definitions of mean and standard deviation for functional data are given, for example, in Ramsay and Silverman [2006]), but it is preferable to operate over the multivariate data, as they are just in the input of the multivariate optimization problem and it is not sure that the changes were preserved in the functional-to-multivariate data transformation step.

2.6 The Algorithm

Let $f_1, \ldots, f_{n_f}$ and $g_1, \ldots, g_{n_g}$ be samples of functions from the two populations, then:

1. **From functional to multivariate data.** For each $f_j,\ j = 1, \ldots, n_f$, the following vector is constructed
$$\mathbf{x}_j^{(f)} = (x_{j1}^{(f)}, \ldots, x_{jp}^{(f)})^t, \tag{56}$$
where $x_{ji}^{(f)} = d(f_j^{i-1)}, \overline{f}^{i-1)}) - d(f_j^{i-1)}, \overline{g}^{i-1)})$. These vectors form the multivariate sample
$$(\mathbf{x}_1^{(f)}, \cdots, \mathbf{x}_{n_f}^{(f)}) = \begin{pmatrix} x_{11}^{(f)} & \cdots & x_{n_f 1}^{(f)} \\ \vdots & \ddots & \vdots \\ x_{1p}^{(f)} & \cdots & x_{n_f p}^{(f)} \end{pmatrix}. \tag{57}$$
The multivariate sample $(\mathbf{x}_1^{(g)}, \cdots, \mathbf{x}_{n_g}^{(g)})$ is defined in an analogous way.

2. **The discriminant function.** These samples are used as input in the optimization problem to obtain the discriminant function:
$$y(\mathbf{x}) = \mathbf{a}^t \mathbf{x}, \tag{58}$$
where $\mathbf{x} = (x_1, \ldots, x_p)^t$, and $\mathbf{a} = \mathbf{a}_F$ or $\mathbf{a} = \mathbf{a}_p$ depending on whether or not the additional constraint is imposed.

3. **The allocation of new curves.** To classify a new curve h, its multivariate vector is constructed:
$$\mathbf{x}^{(h)} = (x_1^{(h)}, \ldots, x_p^{(h)})^t, \tag{59}$$
where $x_i^{(h)} = d(h^{i-1)}, \overline{f}^{i-1)}) - d(h^{i-1)}, \overline{g}^{i-1)})$. Finally, the value $y(\mathbf{x}^{(h)})$ is used to assign the curve h to one of the two populations, as mentioned in subsection 2.4.

Remark 7: As a distance measurement, $d(\cdot, \cdot)$, between two functions we have taken the L_1 distance (and norm) that, for the functions g and f, is defined as:
$$d(f, g) = \|f - g\|_1 = \int_0^1 |f - g|. \tag{60}$$
But other distances can be used following the same approach.

Several versions of this algorithm have been implemented and compared in the following sections.

2.7 Weighted Semidistances or Distances

Let us substitute, for the function h, the discriminant variables into the expression of the discriminant function:

$$\begin{aligned} y(\mathbf{x}^{(h)}) &= \mathbf{a}^t\mathbf{x}^{(h)} = \sum_{i=1}^{p} a_i x_i^{(h)} \\ &= \sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{f}^{i-1)}) - \sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{g}^{i-1)}). \end{aligned} \tag{61}$$

For the linear combinations $\sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{f}^{i-1)})$ and $\sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{g}^{i-1)})$ to take nonnegative values, our additional restrictions ($a_i \geq 0$) is necessary; so only the function $y = \mathbf{a}_p^t\mathbf{x}$ —not the classical linear discriminant function— can be seen as providing a classification based on the minimization of a weighted distance.

As in a space of functions the derivation can imply a loss of information, then

$$\rho(h, \overline{f}) = \sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{f}^{i-1)}) \quad \text{and} \quad \rho(h, \overline{g}) = \sum_{i=1}^{p} a_i d(h^{i-1)}, \overline{g}^{i-1)}), \tag{62}$$

with $a_i \geq 0$ can be interpreted as measurements with a weighted distance if and only if $a_1 \neq 0$; otherwise, it can be interpreted as a measurement with a weighted semidistance, since two functions can differ in a constant and verify $\rho(f, g) = 0$ when $a_1 = 0$. An important general property of $\rho(\cdot, \cdot)$ is that it takes into account at the same time the functions, their smoothness, their curvature, etcetera.

Similarly, when the distance $d(\cdot, \cdot)$ is defined from a norm, that is,

$$d(f, g) = \|f - g\|, \tag{63}$$

the expression (62) can be seen as a weighted norm if, an only if, $a_1 \neq 0$, or as a weighted seminorm otherwise.

3 Simulation Results

In order to illustrate the behavior of our two procedures, we perform a Monte Carlo study using three different settings. In all cases we consider two functional populations in the space $C[0, 1]$ of continuous functions defined in the interval $[0, 1]$. The methods used to classify are the following:

- Distance to the sample functional mean calculated using the functions in the training set (*DFM0*). That is, using the rule

$$\begin{cases} k = 1 & \text{if} \quad x_1 < 0 \\ \\ k = 2 & \text{otherwise} \end{cases} . \tag{64}$$

- Distance to the sample functional mean calculated using the first derivatives of functions in the training set (*DFM1*). That is, using the rule

$$\begin{cases} k = 1 & \text{if} \quad x_2 < 0 \\ \\ k = 2 & \text{otherwise} \end{cases} . \tag{65}$$

- Weighted indicator (*WI*) obtained using our first procedure. Using the algorithm with $\mathbf{x} = (x_1, x_2)^t$ and without the nonnegativity constraint.

- Weighted distance (*WD*) obtained using our second procedure. Using the algorithm with $\mathbf{x} = (x_1, x_2)^t$ and the nonnegativity constraint.

We generate 200 functions from each population. The training set consists of the first 100 functions from each population, and the remaining 100 observations from each sample are the testing set. For each setting we run 1000 replications, so the results are based on 1000 estimates of the misclassifications rates.

Now, we describe the three considered settings.

Simulation setting 1: We consider the following two functional data generating models:

Model B1 $f_i(t) = t + u_i$, where u_i is a uniform random variable on the interval $(0, 1)$.

Model R1 $g_i(t) = t + v_i$, where v_i is a uniform random variable on the interval $(1/2, 3/2)$.

Remark 8: Figure 1(a) displays a random sample for these two models. The sample functional mean for model B1 is marked by circles and for model R1 by squares. Notice that models B1 and R1 differ in level when u_i takes value in $(0, 1/2)$ and v_i in $(1, 3/2)$ but they coincide when u_i and v_i take values in $(1/2, 1)$. This intersection causes a theoretical misclassification rate equal to 25% when the method *DFM0* is used. Moreover, the first derivative of functions f_i and g_i coincides, so method *DFM1* will fail in this setting.

Simulation setting 2: We consider the following two functional data generating models:

Model B2 $f_i(t) = (t + u_i)^2$, where u_i is a uniform random variable on the interval $(0, 1)$.

Model R2 $g_i(t) = t^2 + v_i$, where v_i is a uniform random variable on the interval $(0, 1)$.

Remark 9: Figure 1(b) displays a random sample for these two models. The sample functional mean for model B2 is marked by circles and for model R2 by squares. Notice that models B2 and R2 generate functional observations that cross one another; but if we consider the first derivative, $f_i^{1)}$ and $g_i^{1)}$, then they have significant level differences. The theoretical misclassification rate is equal to 12.5% when the method *DFM1* is used.

Simulation setting 3: We consider the following two functional data generating models:

Model B3 $f_i(t) = (t + u_i)^2 + 5/4$, where u_i is a uniform random variable on the interval $(0, 1)$.

Model R3 $g_i(t) = (t + v_i)^2$, where v_i is a uniform random variable on the interval $(1/2, 3/2)$.

Remark 10: Figure 1(c) displays a random sample for these two models. The sample functional mean for model B3 is marked by circles and for model R3 by squares. Notice that models B3 and R3 also generate functional observations that cross one another (the term $+5/4$ in f is added in order to maximize the crossing) but if we consider the first derivatives, $f_i^{1)}$ and $g_i^{1)}$, then these have level differences in the same way as f_i and g_i generated by models B1 and R1, respectively. So, we have a theoretical misclassification rate equal to 25% when the method *DFM1* is used.

In figure 2 we present the results for the first simulation setting. Figure 2(a) gives the boxplots of the misclassification rates estimates for the four methods. As expected, the method *DFM0* has a misclassification rate of around 25% and the method *DFM1* is useless in this setting. Figures 2(b) and 2(c) give the boxplots of the estimated weights for methods *WI* and *WD*. Both methods give positive weights for the variable associated to f and g and zero weights for the variable associated to $f^{1)}$ and $g^{1)}$. Notice that in this case the variable $Df^{1)} - Dg^{1)}$ has variance equal to zero since $f_i^{1)} = g_i^{1)} = 1$ for all i. In this simulation setting, methods *DFM0*, *WI* and *WD* have the same performance.

In figure 3, we present the results for the second simulation setting. Figure 3(a) gives the boxplots of the misclassification rates estimates for the four methods. In this case, method *DFM0* is outperformed by method *DFM1*, which obtains misclassification rates around the expected 12.5%. Method *WD* has a performance similar to *DFM1*, and both are outperformed by method *WI*. Figures 3(b) and 3(c) give the boxplots of the estimated weights for methods *WI* and *WD*. In this case, method *WI* gives positive weights for the variable associated to f and g and negative

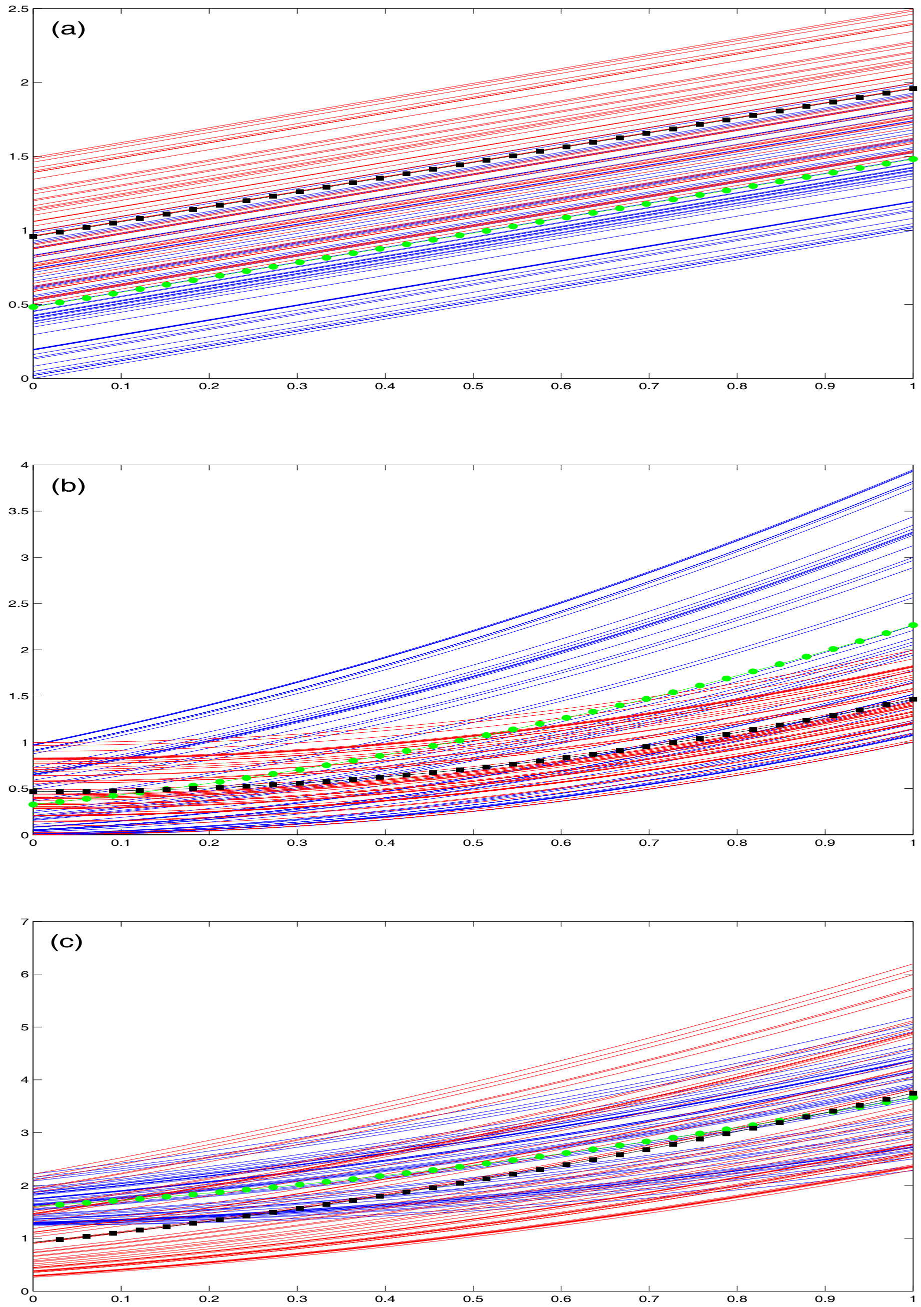

Figure 1: Plots of samples from the three simulation settings: (a) Functions following models B1 and R1; (b) Functions following models B2 and R2; (c) Functions following models B3 and R3.

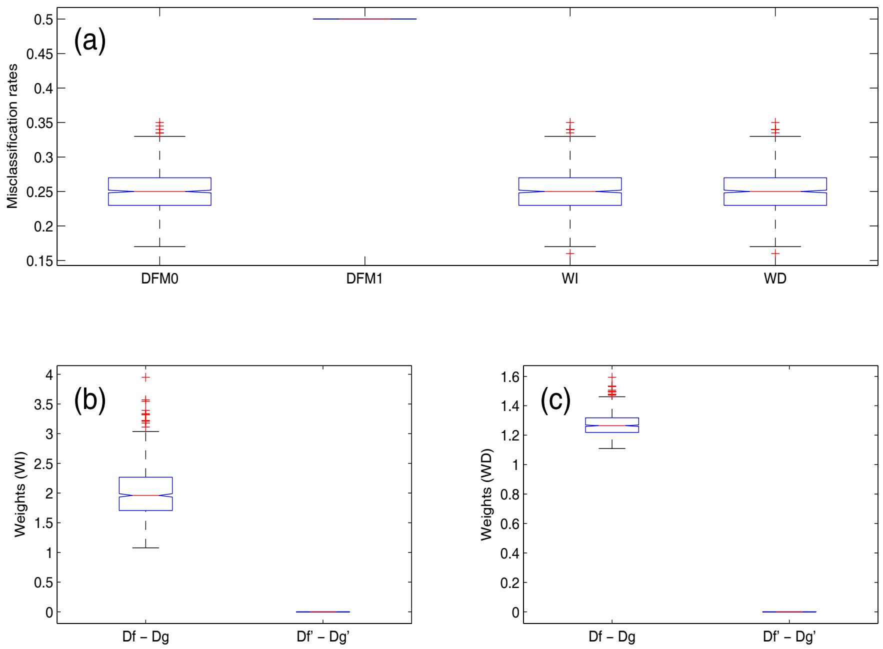

Figure 2: First simulation setting results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

weights for the variable associated to $f^{1)}$ and $g^{1)}$, so the classification rule with *WI* is not a distance. Once we impose the positiveness on the weights, method *WD* gives positive weights for the variable associated to $f^{1)}$ and $g^{1)}$ and zero weights for the variable associated to f and g. So, the classification rule with *WD* is a semidistance. In this setting and in the previous one, method *WD* selects the variable that has lower misclassification rates.

In figure 4, we present the results for the third simulation setting. Figure 4(a) gives the boxplots of the misclassification rates estimates for the four methods. In this case, method *DFM0* is again outperformed by method *DFM1*, which obtains misclassification rates around the expected 25%. Both methods perform worse than the weighted procedures, *WI* and *WD*; method *WI* has the best performance. Here, the improvement comes from the combination of variables associated to functions and their first derivatives. Figures 4(b) and 4(c) give the boxplots of the estimated weights for methods *WI* and *WD*. In this case, method *WI* gives positive weights for the variable associated to f and g in more than 25% of the replications and negative weights in the remaining ones. In all replications, *WI* gives negative weights for the variable associated to $f^{1)}$ and $g^{1)}$. For those replications where there are sign differences, the classification rule with *WI* is not a distance. This “inconvenience” is avoided by using the method *WD*. In this setting, the classification rule with *WD* is a semidistance in all cases and a distance in 75% of the replications.

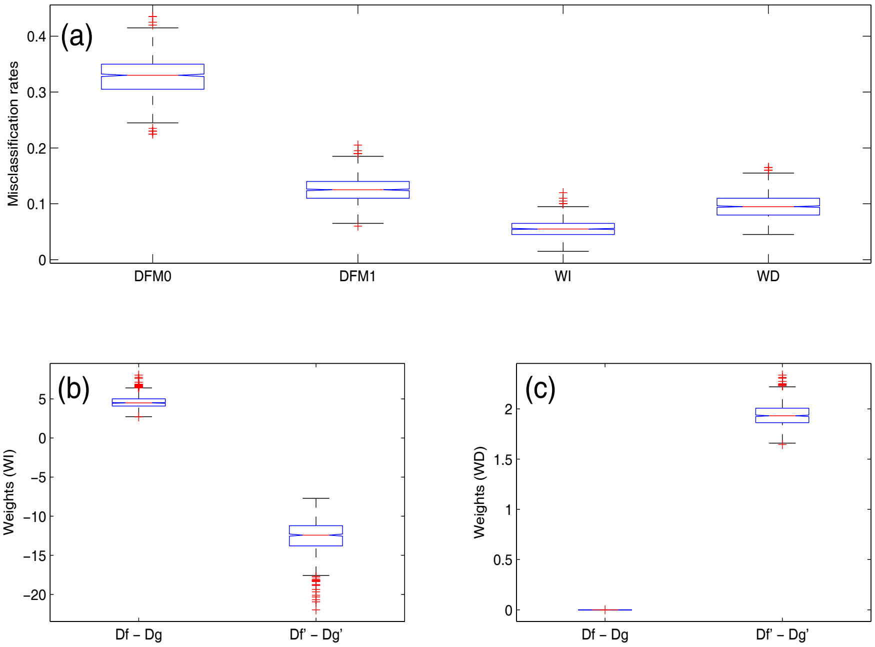

Figure 3: Second simulation setting results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

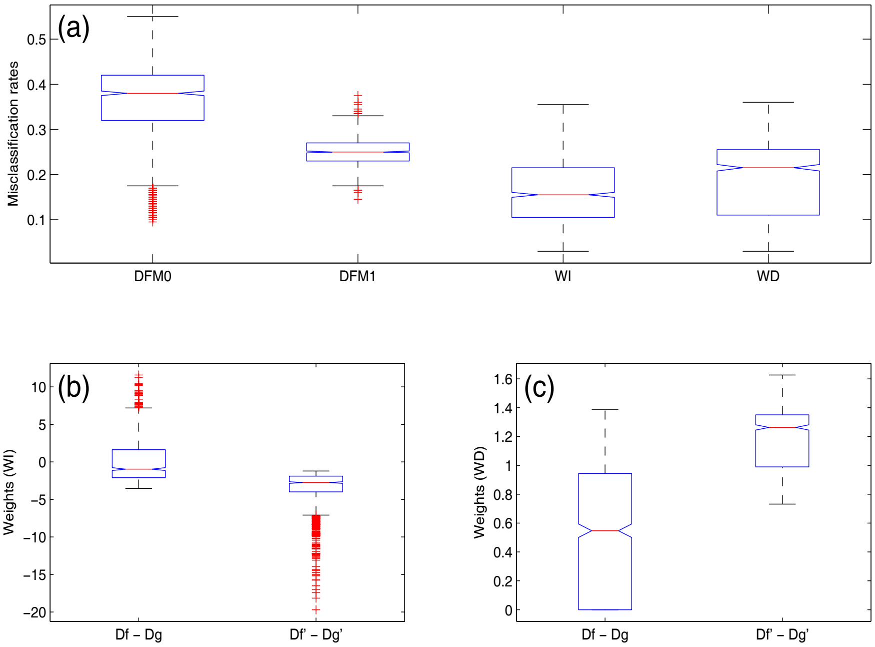

Figure 4: Third simulation setting results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

4 Real Data Examples

In this section we illustrate the performance of our proposal in two benchmark data sets: (**a**) *Spectrometric data set*, consisting of 215 near-infrared spectra of meat samples obtained by a Tecator Infratec Food and Feed Analyzer; (**b**) *Growth curves data set*, consisting of the height (in centimeters) of 44 girls and 39 boys measured at a set of 31 ages from 1 to 18 years old.

In both examples, the original data was smoothed using a cubic smoothing spline with smoothing parameter equal to $1/(1+h^3/6)$, where h is the average spacing of the data sites (see De Boor [1978]).

In this section, the nomenclature for the different versions of the algorithm is that used in the previous section. Furthermore,

- *DFM2* denotes de classification with the distance to the sample functional mean calculated using the second derivatives of functions in the training set. That is, using the rule

$$\begin{cases} k=1 & \text{if} \quad x_3 < 0 \\ & \\ k=2 & \text{otherwise} \end{cases} . \tag{66}$$

- Now the weighted approaches consider up to the second derivative by taking $\mathbf{x} = (x_1, x_2, x_3)^t$.

4.1 Spectrometric Data

The classification problem in the spectrometric data set consists in separating meat samples with a high fat content (more than 20%) from samples with low fat content (less than 20%). Among the 215 samples, 77 have high fat content and 138 have low fat content. Figure 5 shows a sample of these 100-channel absorbance spectrum in the wavelength 850–1050 nm and the first and second derivatives.

Among others, Ferraty and Vieu (2003), Rossi and Villa (2006) and Li and Yu (2008) had considered the original spectrum and its derivatives for classification purpose and had concluded that the second derivative produces the lower misclassification rates.

In order to evaluate the performance of our proposal, we will split the data set into 120 spectra for training and 95 spectra for testing as in Rossi and Villa (2006) and Li and Yu (2008). The classification results shown in figure 6 are based on 1000 replications. Methods *WI* and *WD* obtain a mean misclassification rate equal to 2.02% and 2.32%, respectively. They improve the classification rule based on the second derivative, *DFM2*, which obtains 3.70%.

In this example, method *WI* gives positive weights to the variable associated with f and g, and negative weights for the variables associated with $f^{1)}$ and $g^{1)}$ and with $f^{2)}$ and $g^{2)}$; so the classification rule with *WI* is not a distance. Method *WD* gives positive weights to the variables associated with $f^{1)}$ and $g^{1)}$ and with $f^{2)}$ and $g^{2)}$, and zero weights for the variable associated with

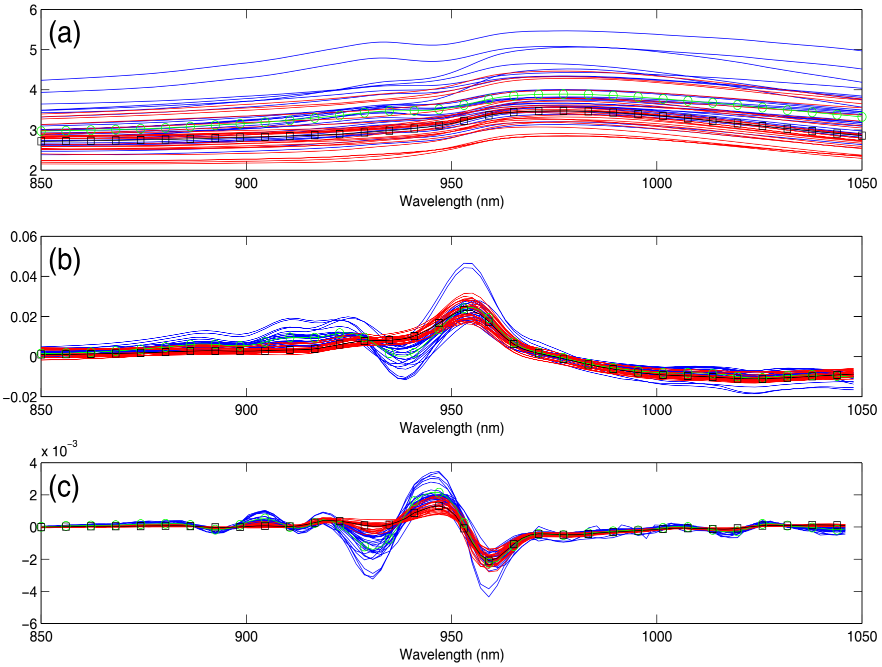

Figure 5: Sample from the spectrometric data set (wavelengths 850–1050 nm): (a) Spectrum; (b) First derivative; (c) Second derivative.

f and g. Notice that both procedures give the higher weights to the variable associated with $f^{2)}$ and $g^{2)}$, which is consistent with the results of Ferraty and Vieu (2003), Rossi and Villa (2006) and Li and Yu (2008).

The functional support vector machine proposed by Rossi and Villa (2006) obtains 3.28% (7.5%) using a linear (Gaussian) kernel and the spectra, and a 2.6% (3.28%) using a Gaussian (linear) kernel and the second derivative of the spectra.

The nonparametric functional method proposed by Ferraty and Vieu (2003) obtains a mean error of around 2% using the second derivative. Notice that Ferraty and Vieu (2003) use a training set with 160 spectra. In that setting, our mean misclassification rates are equal to 1.89% and 2.27%, respectively.

Li and Yu (2008) obtain 3.98%, 2.91% and 1.09% using the raw data, the first derivative and the second derivative, respectively. Notice that Li and Yu's method selects the data segments where the two populations have large differences, and then it combines the linear discriminant as a data reduction tool and the support vector machine as classifier. Li and Yu's method has three tuning parameters — number of segments, the separation among segments and the regularization parameter of the support vector machine.

If we repeat our procedures using the channels in the wavelengths 1000–1050 nm, then we obtain 1.49% and 1.30%, using *WI* and *WD*, respectively. Figure 7 shows a sample of these spectrum in the wavelength 1000–1050 nm and the first and second derivatives. This segment, 1000–1050 nm, was obtained by cross-validation through a grid search. The design of a segmenta-

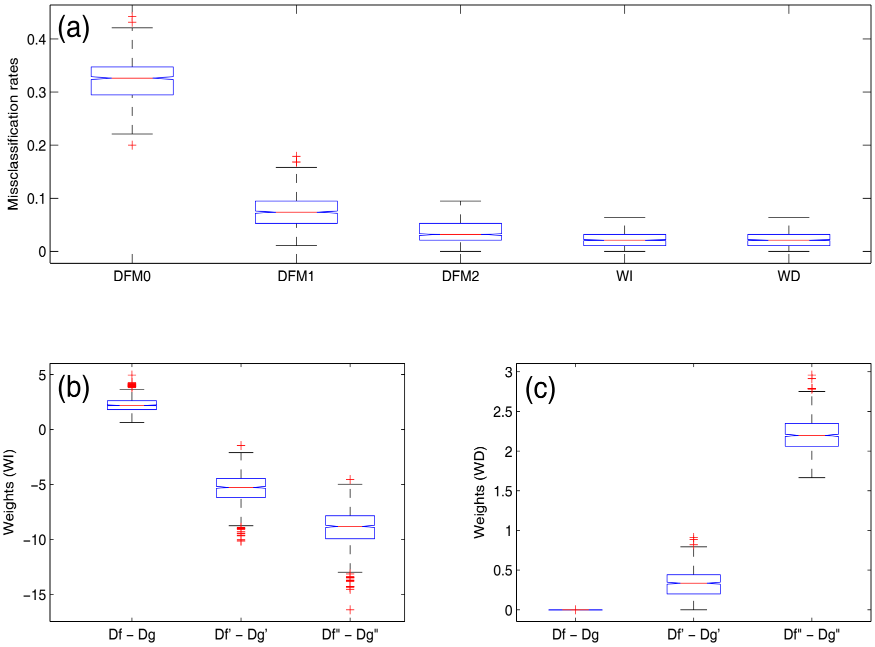

Figure 6: Spectrometric data set classification results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

tion approach for selecting more than one segment is beyond the scope of this paper and probably deserves separate research.

4.2 Growth Data

The classification problem in the growth data set consist of separating samples by sex, taking the growth curves as variables. Figure 8 shows a sample of these curves, measured in ages ranging from in $[1, 18]$, and their first and second derivatives. López-Pintado and Romo (2006) had considered the growth curves (but not their derivatives) for classification purpose.

In order to evaluate the performance of our proposal, we will split the data set into 60 curves for training and the remaining 33 for testing. The classification results shown in figure 9 are based on 1000 replications. Weighted methods *WI* and *WD* have similar behavior, with means misclassification rates equal to 3.65% and 3.75%, respectively. They improve the classification rules based on the raw data, on the first derivative or on the second derivative, which obtain 31.08%, 5.30% and 18.85%, respectively. The best result with the depth-based classification procedure proposed by López-Pintado and Romo (2006) was 14.86%.

In this example, method *WI* gives positive weights for the variable associated to $f^{2)}$ and $g^{2)}$ and negative weights for the variables associated to $f^{1)}$ and $g^{1)}$; so the classification rule with *WI* is not a distance. Method *WD* gives positive weights for the variables associated to f and g and

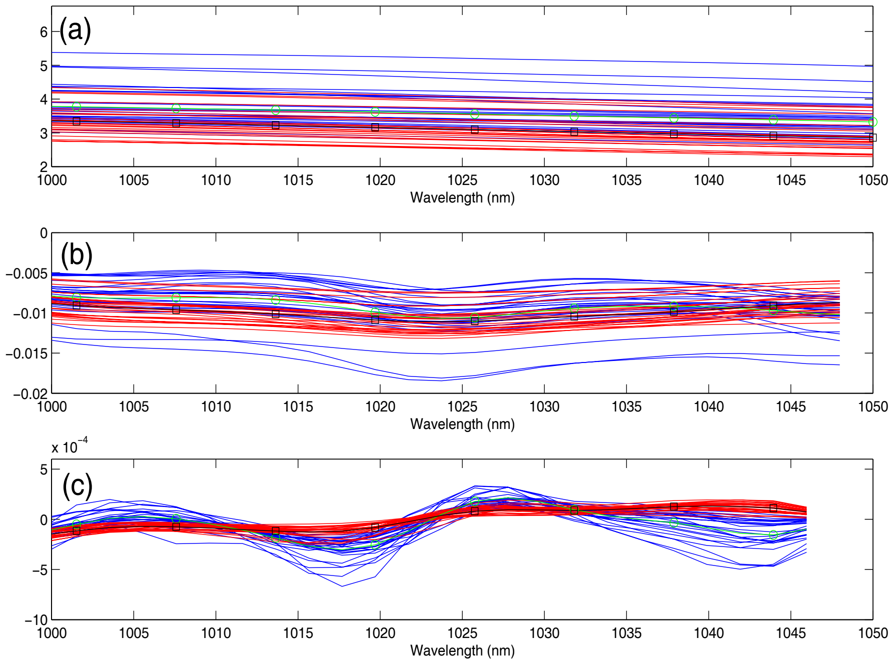

Figure 7: Sample from the spectrometric data set (wavelengths 1000–1050 nm): (a) Spectrum; (b) First derivative; (c) Second derivative.

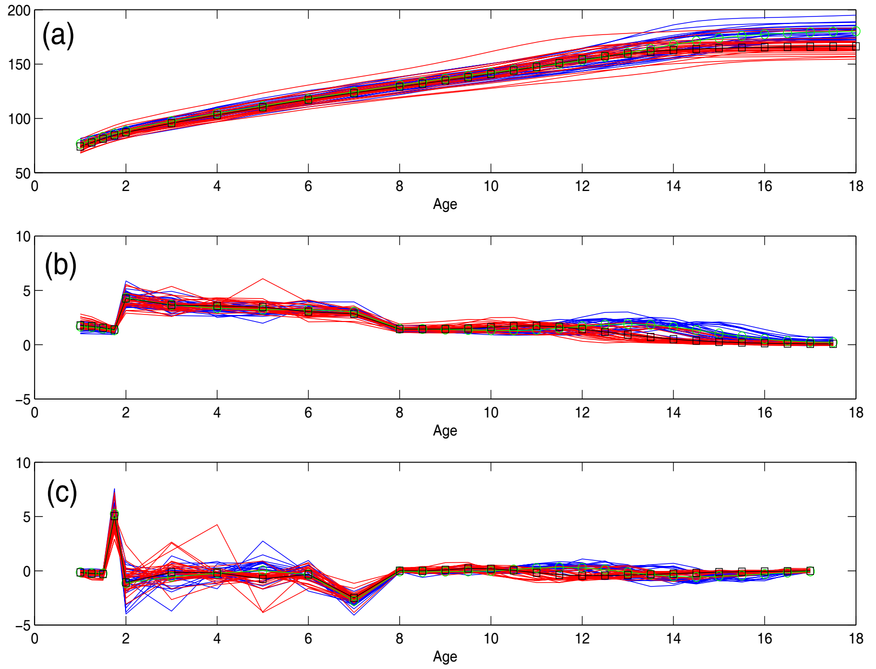

Figure 8: Sample from the growth data set: (a) Spectrum; (b) First derivative; (c) Second derivative.

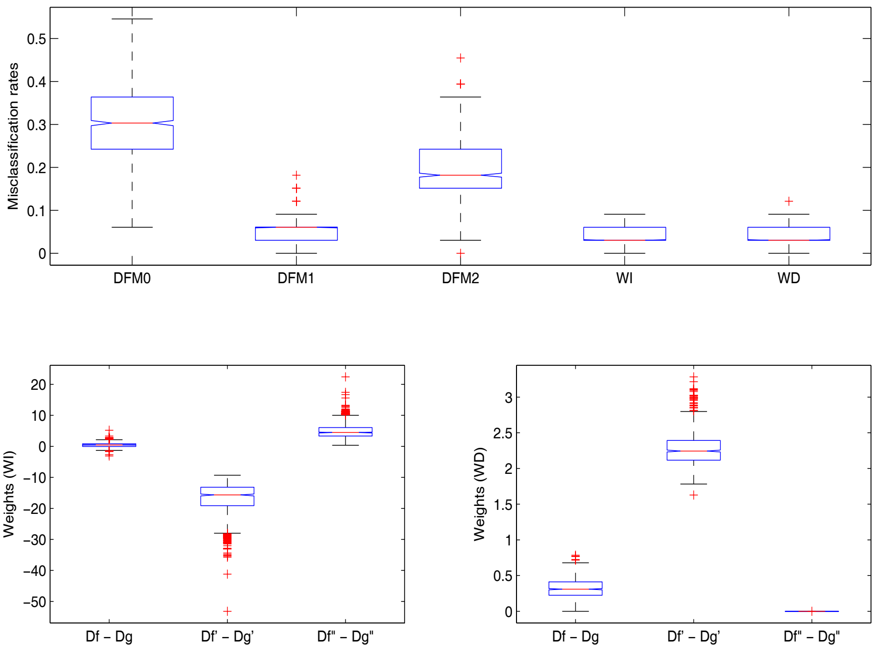

Figure 9: Growth data set classification results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

to $f^{1)}$ and $g^{1)}$; then, the classification rule is a distance.

5 Conclusions

In this paper we have proposed a new approach for discriminating functional data. This method involves the use of distances to a representative function and its successive derivatives. Our simulation studies and our applications show that the method performs very well resulting in small training and testing classification errors. Applications to real data show that our procedure performs as well as —and in some cases better than— other classifications methods. In addition, our methodology provides, through the weights, information about the importance of each data transformation. Finally, some adaptability of our methodology to the different types of functional data can be achieved by selecting the distance $d(\cdot,\cdot)$ or the multivariate classification technique.

Acknowledgements

This research has been supported by CICYT (Spain) grants SEJ2007-64500 and ECO2008-05080. The first author also acknowledges the support of "Comunidad de Madrid" grant CCG07-UC3M/ HUM-3260.

References

[1] Abraham, C., P.A. Cornillon, E. Matzner-Løber and N. Molinari (2003). Unsupervised Curve Clustering using B-Splines. *Scandinavian Journal of Statistics*. 30 (3), 581–595.

[2] Abraham, C., G. Biau and B. Cadre (2006). On the Kernel Rule for Function Classification. *AISM*. 58, 619–633.

[3] Baíllo, A., and A. Cuevas (2008). Supervised Functional Classification: A Theoretical Remark and Some Comparisons. *Manuscript available at http://arxiv.org/abs/0806.2831*.

[4] Berlinet, A., G. Biau and L. Rouvière (2008). Functional Supervised Classification with Wavelets. *Annales de l'I.S.U.P.*. 52 (1–2), 61–80.

[5] Biau, G., F. Bunea and M.H. Wegkamp (2003). Functional Classification in Hilbert Spaces. *IEEE Transactions on Information Theory*. 1 (11), 1–8.

[6] Boor, C. de (1978). *A Practical Guide to Splines*. Springer-Verlag

[7] Cuesta-Albertos, J.A., and R. Fraiman (2007). Impartial Trimmed k-Means for Functional Data. *Computational Statistics and Data Analysis*. 51, 4864–4877.

[8] Dabo-Niang, S., F. Ferraty and P. Vieu (2006). Mode Estimation for Functional Random Variable and Its Application for Curves Classification. *Far East Journal of Theoretical Statistics*. 18 (1), 93–119.

[9] Dabo-Niang, S., F. Ferraty and P. Vieu (2007). On the Using of Modal Curves for Radar Waveforms Classification. *Computational Statistics and Data Analysis*. 51, 4878–4890.

[10] Ferraty, F., and P. Vieu (2003). Curves Discrimination: A Nonparametric Functional Approach. *Computational Statistics and Data Analysis*. 44, 161–173.

[11] Ferraty, F., and P. Vieu (2006). *Nonparametric Functional Data Analysis*. Springer.

[12] Hall, P., D.S. Poskitt and B. Presnell (2001). A Functional Data-Analytic Approach to Signal Discrimination. *Technometrics*. 43 (1), 1–9.

[13] Hastie, T., A. Buja and R.J. Tibshirani (1995). Penalized Discriminant Analysis. *The Annals of Statistics*. 23 (1), 73–102.

[14] Jagannathan, R., and T. Ma (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraint Helps. *The Journal of Finance*. 58 (4), 1651–1683.

[15] James, G.M., and T. Hastie (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society*. Series B, 63, 533–550.

[16] James, G.M., and C. A. Sugar (2003). Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association.* 98 (462), 397–408.

[17] Kiers, H.A.L. (1995). Maximization of Sums of Quotients of Quadratic Forms and Some Generalizations. *Psychometrika.* 60 (2), 221–245.

[18] Li, B., and Q. Yu (2008). Classification of Functional Data: A Segmentation Approach. *Computational Statistics and Data Analysis.* 52, 4790–4800.

[19] López-Pintado, S., and J. Romo (2006). Depth-Based Classification for Functional Data. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications.* American Mathematical Society. DIMACS Series, 72, 103–121. (R. Liu, R. Serfling and D.L. Souvaine [eds]).

[20] McDonald, R.P. (1979). Some Results on Proper Eigenvalues and Eigenvectors with Applications to Scaling. *Psychometrika.* 44 (2), 211–227.

[21] Nerini, D., and B. Ghattas (2007). Classifying Densities Using Functional Regression Trees: Applications in Oceanology. *Computational Statistics and Data Analysis.* 51, 4984–4993.

[22] Ramsay, J.O., and B.W. Silverman (2006). *Functional Data Analysis.* Springer.

[23] Rossi, F., and N. Villa (2006). Support Vector Machine for Functional Data Classification. *Neurocomputing.* 69, 730–742.

[24] Tarpey, T., and K.K.J. Kinateder (2003). Clustering Functional Data. *Journal of classification.* 20 (1), 93–114.

[25] Wang, K., and T. Gasser (1997). Alignment of Curves by Dynamic Time Warping. *The Annals of Statistics.* 25 (3), 1251–1276.

[26] Wang, K., and T. Gasser (1999). Synchronizing Sample Curves Non-Parametrically. *The Annals of Statistics.* 27 (2), 439–460.