Sapienza, University of Rome
Department of Statistical Sciences

Ph.D. Degree in Methodological Statistics

Statistical Methodology for Classification of Partially Observed Functional Data

Candidate:
Marco Stefanucci

Thesis advisor:
Prof. Pierpaolo Brutti

Thesis submitted in 2018

Abstract

This thesis is devoted to the development of new methodologies for the classification of partially observed functional data. Functional Data Analysis is nowadays one of the most active area of research in statistics. It deals mostly with data coming from technical machineries and digital instruments, treating data as functions. Classification of this kind of data is still an open problem and there are several available methods in the literature. Unfortunately, none of these methods is directly applicable when the data are partially observed, i.e. exhibit some missing parts. The aim of this work is to provide new insights and proposals for discrimination of functional fragments. The theory we develop is strongly supported by extensive simulations and all the methods are illustrated on a real medical dataset called Aneurisk, on which we outperform previous classification performance.

Contents

Chapter 1

Introduction

1.1 What is Functional Data Analysis?

1.1.1 A brief introduction

Functional Data Analysis is a field of statistics concerning the statistical analysis of samples of functions. The first thing that comes to the mind when one think about functions is a curve. However, functional data analysis deals also with more *complex* objects, as surfaces and, in general, functions of n variables. Functional Data Analysis (FDA) has expanded rapidly in the last three decades. The exponential growth of this field can be explained by the continuous interplay between the development of pure theoretical concepts and models on one side, and the wide spectrum of real data applications on the other side.

From the late 80s, researchers in statistics started to think about functions as observations or units of interest, and the first theoretical contributions to the field were collected in the mid 90s on the well known book by Ramsay and Silverman (Ramsay and Silverman (2005)) . With this work as the state of the art, several progresses were made until now, translating the multivariate statistics paradigm into the functional one as well as developing *ex novo* methodologies. Beyond the book by Ramsay and Silverman, relevant monographs on the field are Horváth and Kokoszka (2012), Hsing and Eubank (2015) and Kokoszka and Reimherr (2017). Nowadays, most of the standard statistics theory for functional data has been developed and the focus of the researchers has moved towards more complex problems. One of these problems arises when it is not possible to observe each function

completely over the domain, but only *partially*, on a subset of it. Under this framework it is not straightforward to implement any of the existing methodologies, and necessary adjustments must be done.

From the practical point of view, FDA continues to find a lot of applications. This is mostly due to the increasing diffusion in the society of technologies able to register data almost continuously, like for example smartwatches, termometers or medical machineries. Data arising from this kind of measurements can be easily thought as realizations of an underlying continuous function and then being analyzed by tools coming from FDA. As an example, the dataset that we will analyze in the subsequent chapters comes from a medical study and in particular from images obtained by three dimensional angiographies. The main feature of these data is the fact that they are partially observed, in the sense explained above. We will show how to treat such data, and we will develop specific methodology in the context of supervised classification (Izenman (2009)).

The rest of the chapter is devoted to the introduction of the main tools of FDA, to a brief survey of classification methods and to a more detailed introduction to the subject of partially observed functional data.

1.1.2 Basis representation

We generally put ourselves into the framework in which we have a sample of N independent *univariate* random functions (e.g. curves) $z_1(t), z_2(t), \ldots, z_N(t)$, realization of the smooth process $Z(t)$. Here t stands for time, but it can be any index of continuity, depending on the problem. In general, the curves are observed on a finite grid of points and with error. We denote with $T_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})$ the vector of time points where the $i^{\texttt{th}}$ curve is observed and we write

$$x_i(t_{ij}) = z_i(t_{ij}) + \varepsilon_{ij}. \tag{1.1}$$

Thus the observed curve on T_i can be thought as a realization of the underlying continuous process $Z(t)$ sampled on T_i plus a measurement error term. The case where $T_i = T$ for each i is known as regular design. In this case the data can be stored in a matrix and standard statistical procedures can be applied without any further preprocessing.

The smoothness of the process $Z(t)$ is a common assumption. The realizations $z_i(t), z_2(t), \ldots, z_N(t)$ should be smooth too, but we are only able to observe the process with error. If this error is not negligible, we need to denoise (i.e. remove the high-frequency oscillations) the observed data to get smooth trajectories. This problem, known in the literature as *smoothing*, can be approached by different points of view but since the focus of this work is somewhat different we are not going to present all the approaches here, we just limit our attention to the basis expansion approach.

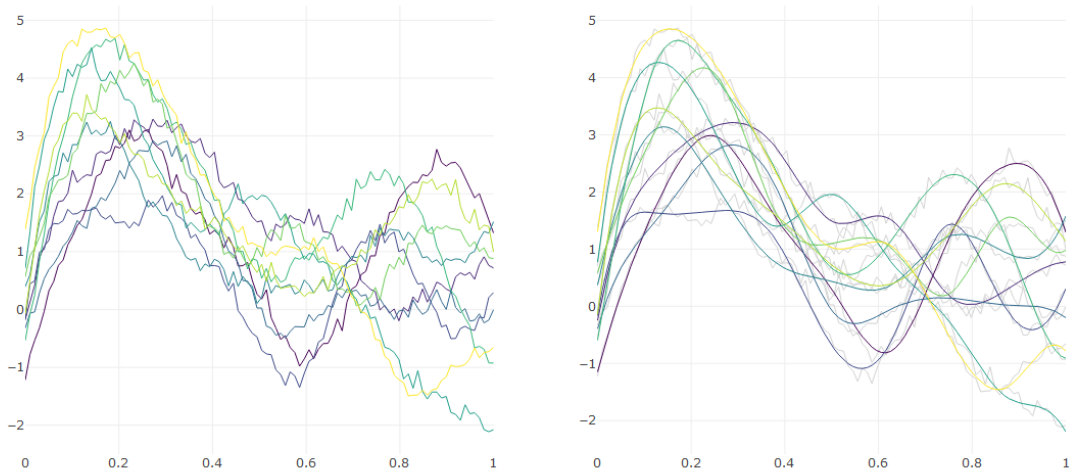

Figure 1.1: *Basis representation for synthetic data. Left: Ten curves generated from a gaussian process observed with error on a grid of 100 time points in [0,1]. Right: Same set of curves represented using a B-spline basis of 10 elements.*

Let $L_2(0,1)$ denote the set of functions $f : [0,1] \to \mathbb{R}$ such that $||f|| = \int_0^1 f^2(t)\mathrm{d}t < \infty$. A sequence of functions $\eta_1, \eta_2, \ldots$ is called *orthonormal* if $||\eta_k|| = 1$ for each k and $\langle \eta_k, \eta_m \rangle = \int_0^1 \eta_k(t)\eta_m(t)\mathrm{d}t = 0$ for $k \neq m$. A sequence is complete if the only function that is orthogonal to each η_k is the zero function. A complete, orthonormal, set of functions forms a basis meaning that if $f \in L_2(0,1)$ then f can be expanded as

$$f(t) = \sum_{k=1}^{\infty} \theta_k \eta_k(t) \tag{1.2}$$

where $\theta_k = \langle f, \eta_k \rangle = \int_0^1 f(t)\eta_k(t)\mathrm{d}t$. From the practical point of view, one can use basis expansion to represent the functional datum and to filter out some undesirable features of the observed curves. For this purpose consider the truncate series

$$x_i(t) \simeq \sum_{k=1}^{K} \theta_{ik}\eta_k(t) \tag{1.3}$$

for some scalars $\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}$ and basis functions $\eta_1(t), \eta_2(t), \ldots, \eta_K(t)$. If the basis functions are chosen to be smooth, this smoothness will be reflected in the representation of the functions $x_1(t), x_2(t), \ldots, x_N(t)$. The most common choices for the basis are the Fourier series and the splines. In particular, even if they are not orthonormal, the use of splines to represent and analyze functional data has became quite common in the applications. Using such a truncate series allow to retain the main features of the curves while ignoring the high-frequency fluctuations. This is crucial because the process $Z(t)$ is supposed to be smooth.

Once the observations are represented in a way determined by the researcher, it is useful to plot the sample in order to see the behaviour of the curves. The graphical inspection is more important than one might think and it is quite powerful with respect to the case of multivariate data. An illustration of the smoothing procedure can be seen in figure 1.1.

1.1.3 Sample mean and covariance

The concepts of sample mean and covariance are not really different compared to the case of multivariate analysis. Starting from smoothed trajectories, the sample pointwise mean will result in a smooth curve, representing the average behaviour of the sample. It is defined as

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t). \tag{1.4}$$

The sample pointwise variance gives us an idea about the typical variability of curves at any point t and it is defined in a similar way as

$$\widehat{\sigma}^2(t) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i(t) - \bar{x}(t))^2. \tag{1.5}$$

From the computational point of view, one must evaluate (1.4) and (1.5) on each time point of the grid T. The formula (1.5) gives no information on how the values of the curves at point t relates to those at point s. An object which is extensively used in FDA is the sample covariance function defined as

$$\widehat{\rho}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)). \tag{1.6}$$

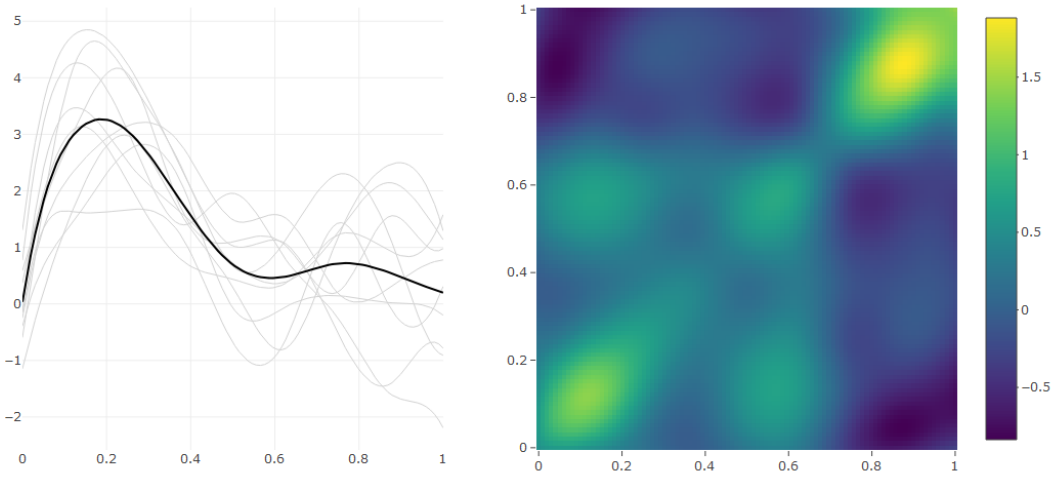

Figure 1.2: *Sample mean and covariance for synthetic data. Left: sample mean of smoothed curves. Right: covariance surface of smoothed curves.*

The interpretation of the values of $\widehat{\rho}(s,t)$ is the same as for the usual variance-covariance matrix. for example, large values indicate that $x_i(s)$ and $x_i(t)$ tend to be simultaneously above or below the average values at these points. In the case of (1.6) one must evaluate the function on each possible couple $\{s,t\}$ in $T \times T$. An example of sample mean and covariance can be found in figure 1.2.

If we denote by $\mu(t) = \mathbb{E}Z(t)$ and $\rho(s,t) = \mathbb{E}[(Z(s) - \mu(s))(Z(t) - \mu(t))]$ the true mean and covariance functions of the process, we can think of (1.4) and (1.6)

as estimators of the population parameters $\mu(t)$ and $\rho(s,t)$. It can be shown that these estimators are consistent (Horváth and Kokoszka (2012)).

1.1.4 Functional Principal Component Analysis

Eigenanalysis plays a central role in functional analysis. In a general vector space, suppose $\mathscr{L}$ is an operator and λ a real number. The number λ is called an *eigenvalue* of $\mathscr{L}$ if there is a nonzero vector v such that

$$\mathscr{L}v = \lambda v \tag{1.7}$$

Every such vector v is called an *eigenvector* of $\mathscr{L}$ corresponding to λ. If the operator acts on a function space, v is often called an *eigenfunction*. We now list some basic facts that are of fundamental importance throughout this work.

- An operator $\mathscr{L}$ over a function space is called *nonnegative-definite* if for every function f we have $\langle \mathscr{L}f, f\rangle \geq 0$. All eigenvalues of a nonnegative-definite operator are nonnegative.

- An operator $\mathscr{L}$ over a function space is called *symmetric* or *self-adjoint* if for any f and g we have $\langle \mathscr{L}f, g\rangle = \langle f, \mathscr{L}g\rangle$. Eigenfunctions corresponding to distinct eigenvalues of a symmetric operator are orthogonal.

- Given the orthonormal system $\{f_k, k \geq 1\}$, consisting of eigenfunctions of a symmetric Hilbert-Schmidt operator $\mathscr{L}$ corresponding to nonzero eigenvalues λ_k, every z in the function space where $\mathscr{L}$ acts has a unique representation
$$z = \sum_{k=1}^{\infty} a_k f_k + h,$$
where h satisfies $\mathscr{L}h = 0$.

These statements are telling us that, given a symmetric, non-negative definite operator on a function space, each function in that space can be written as an infinite linear combination of eigenfunctions of the operator.

We have already introduced the covariance function $\rho(s,t)$ as the covariance between two distinct points of the process $Z(t)$. Now denote by $\mathscr{R}$ the integral operator defined as

$$\mathscr{R}f(\cdot) = \int \rho(\cdot,t)f(t)\mathrm{d}t. \tag{1.8}$$

This operator is called the *covariance operator*: it is symmetric and nonnegative-definite, so the statements apply. We write the eigenproblem related to $\mathscr{R}$ as

$$\mathscr{R}\varphi = \lambda\varphi, \tag{1.9}$$

so $\{\varphi_k, k \geq 1\}$ are the eigenfunctions – or principal components – of $\mathscr{R}$ corresponding to the eigenvalues $\{\lambda_k, k \geq 1\}$, usually listed in decreasing order. The third statement permits us to write the expansion of a random function in $L_2(0,1)$:

$$z(t) - \mu(t) = \sum_{k=1}^{\infty} \xi_k \varphi_k, \tag{1.10}$$

where $\xi_k = \langle z, \varphi_k \rangle = \int z(t)\varphi_k(t)\mathrm{d}t$ are called *scores*. In practice, one observe the functions $z_1(t), z_2(t), \ldots, z_N(t)$ with error on a finite grid and needs also to empirically estimate the eigenfunctions. Thus the final representation of the data will result

$$x_i(t_{ij}) - \widehat{\mu}(t_{ij}) = \sum_{k=1}^{K} \widehat{\xi}_{ik}\widehat{\varphi}_k(t_{ij}) + \varepsilon_{ij}, \tag{1.11}$$

where quantities denoted with the hat are sample counterparts of population quantities. The series is truncated to the first K components because it is not possible to estimate an infinite number of functions from the data. Moreover, in general a small value of K is enough to represent the data in an optimal way, in the sense that will be clarified in the next paragraph. We refer to (1.11) as the principal component representation of the observed data. An example of fPCA is shown in figure 1.3.

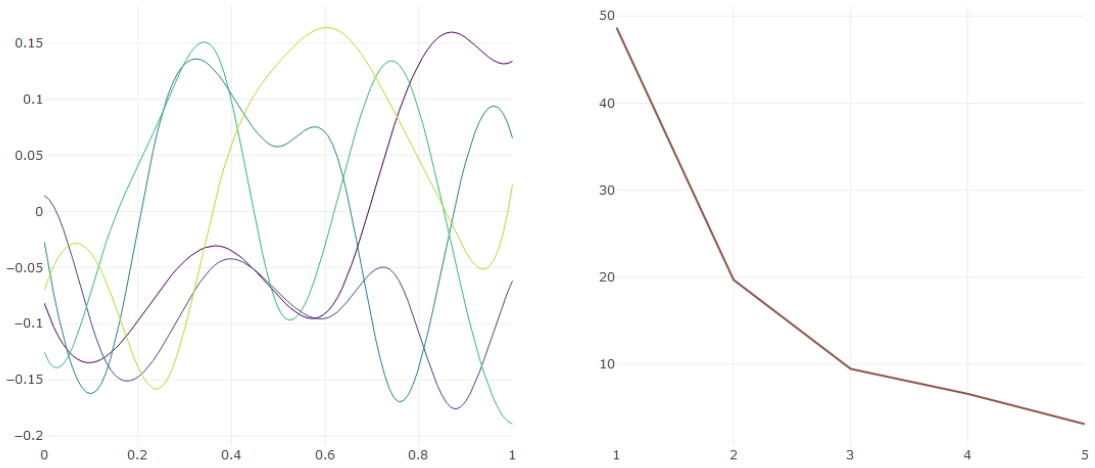

Figure 1.3: *Functional Principal Component Analysis (fPCA) for synthetic data. Left: first 5 eigenfunctions of the sample covariance operator. Right: scree-plot of the first 5 eigenvalues.*

As we have seen in section 1.1.2, it is of fundamental importance in FDA to expand the observed functions $x_i(t)$ as $\sum_{k=1}^{K} \theta_{ik}\eta_k(t)$, where the $\eta_k(t)$ are some basis functions. We now give a justification for the use of the empirical principal component functions as a basis to represent the data.

The projection of a random function $z(t)$ on the subspace spanned by $\eta_1(t), \eta_2(t), \ldots, \eta_K(t)$ is $\sum_{k=1}^{K}\langle z, \eta_k\rangle\eta_k(t)$. We say that the orthonormal system $\eta_1(t), \eta_2(t), \ldots, \eta_K(t)$ is optimal in the sense of minimizing the expected distance between z and its span if minimizes

$$S(\eta_1, \eta_2, \ldots, \eta_K) = \mathbb{E}\left\| z(t) - \sum_{k=1}^{K}\langle z, \eta_k\rangle\eta_k(t)\right\|^2 \tag{1.12}$$

over all orthonormal systems of K functions. It can be shown that (1.12) is minimized when, for each k, η_k is equal to φ_k, the k^{th} eigenfunction of the covariance operator $\mathscr{R}$ of the process $Z(t)$. Thus using (sample) eigenfunctions of $\mathscr{R}$ as a basis to represent the data leads to a low dimensional representation that maximizes the explained (sample) variance.

1.2 Classification of functional data

1.2.1 A brief survey

In this section we recall the framework of supervised classification in the context of functional data and review some of the existing methods to approach this problem.

Suppose that a random function $z(t)$ can be generated either by population 0, which has $\mu_0(t)$ as the true mean function of the process and $\rho_0(s,t)$ as the true covariance function, or population 1, which has $\mu_1(t)$ as the true mean and $\rho_1(s,t)$ as the true covariance.

The available data consist on a sample of N independent observations $\{x_i(t), y_i\}$, $i = 1, \ldots, N$ where $x_i(t)$ is a curve observed with error and $y_i \in \{0,1\}$ is a binary variable representing the population membership. The problem is to decide, from the information provided by the training sample, if a new observation, $x_{\texttt{new}}(t)$ – for which the corresponding $y_{\texttt{new}}$ is not known – has been taken in population 0 or 1. The mathematical aim is to find a "rule" or "classifier" $g : L_2(0,1) \to \{0,1\}$ that minimizes the misclassification error, defined as $\mathbb{P}(g(x) \neq y)$. The optimal classification rule – sometimes called *Bayes rule* – is usually unknown, but it can be approximated in different way from the sample.

Extensions of classical multivariate methods for classification in the context of functional data has been developed. Ferraty and Vieu (2003) proposed a non-parametric kernel estimation of the classification rule g while Biau et al. (2005) proved the consistency of the k-nn classifier (Izenman (2009)) for functional data. Important contributions are the works by Rossi and Villa (2006) and Shin (2008) which translated in a functional context the support vector machines (Izenman (2009)) and linear discriminant analysis paradigms, respectively.

A successful approach to the classification task, largely used in the applications, is to represent the infinite–dimensional curves via basis expansion. Since, empirically, the series (1.3) must be finite, the original data are projected into a finite–dimensional space and standard multivariate techniques can be employed. Works towards this direction can be classified depending on the kind of basis used to represent the data and on the classification rule applied after the projection. One of the first contribution was the work by Hall et al. (2001) which combined a fPCA–based dimensionality reduction with a discriminant analysis in the space of

the scores. Specifically, the authors projected the data into the principal component basis (see section 1.1.4) and then applied a quadratic discriminant analysis Izenman (2009) using the scores computed by (1.11). A similar research was conducted by Glendinning and Herbert (2003), including a smoothing step in the procedure. Leng and Müller (2006) and Song et al. (2008) further worked with fPCA representation but changed the classification rule. Example of use of different basis in the representation step are, for example, Berlinet et al. (2008) and Wang et al. (2007) which considered wavelets in the costruction of a classification model, from both frequentist and bayesian point of view.

One potential problem with methods based on projections is that, truncating the series, one may miss some features on the data that relates the curves to the response. Using bases like Fourier, splines, wavelets and also the empirical principal components there is no guarantee of retain a great amount of the relation between X and Y. As we have seen in section 1.1.4, the principal components basis is optimal in the sense of explained variance and this quantity has no relation with the response. A possible way to overcome this problem is to consider a basis that takes into account the *covariation* between X and Y, as for example the Partial Least Squares (PLS) basis. The algorithm related to PLS solution of a linear problem will be discussed in detail in the next chapter, together with its thoretical properties. For the moment, it is important to know that the PLS basis is optimal in the sense of explained *covariance*. In the context of classification, Preda et al. (2007) introduced a PLS–based procedure starting from the functional linear model (Cardot et al. (1999)). Escabias et al. (2007) further developed a similar idea with the functional logistic regression (Müller and Stadtmüller (2005)). The works by Delaigle and Hall (2012a) and Delaigle and Hall (2012b) elucidated the behaviour of PLS estimation from a theoretical point of view. Nevertheless, from a practical point of view it is not clear which basis performs better. A recent comparison can be found in Febrero-Bande et al. (2017).

Lastly, it worth to mention another approach used in classification: it is based on the idea that, starting from the observed curves, only few time points of T are useful for the classification task. Under this framework, the works by Ferraty et al. (2010), Delaigle et al. (2012) and Berrendero et al. (2016) are mainly devoted to the optimal selection of the most relevant points.

The work by Delaigle and Hall (2012a) constituites a fundamental discovery in the field of classification of functional data. In simple words, even though curves exhibit an infinite–dimensional nature, the problem of classification is often simpler with respect to the multivariate case and in specific situations the misclassification error can be zero, resulting in the so called *perfect classification phenomenon*. Since this fact is of high relevance on our work, this phenomenon will be the focus of the next section.

1.2.2 The perfect classification phenomenon

The term *perfect classification* refers to the possibility of a zero misclassification error that can occur in classifying functional data. Delaigle and Hall (2012a) face the problem considering the so called *centroid classifier*, which is based on the distance between the new observation $x_{\texttt{new}}(t)$ and both the sample means of the two groups, $\bar{x}_0(t)$ and $\bar{x}_1(t)$. A generic distance is denoted by $D(\cdot,\cdot)$ and the authors propose the distance $D(f,g) = |\langle f,\psi\rangle - \langle g,\psi\rangle|$, where the function $\psi(t)$ is called the projecting function. The decision rule, $\mathbf{T}(x_{\texttt{new}})$ is given by

$$\mathbf{T}(x_{\texttt{new}}) = D^2(x_{\texttt{new}},\bar{x}_0) - D^2(x_{\texttt{new}},\bar{x}_1). \tag{1.13}$$

The centroid classifier assigns $x_{\texttt{new}}$ to population 1 or 0 according to whether the statistic $\mathbf{T}(x_{\texttt{new}})$ is positive or negative, respectively.

The function $\psi(t)$ is chosen to minimize the classification error. Under the assumption of gaussianity and of equal covariance kernel among populations, $\rho_0(s,t) = \rho_1(s,t) = \rho(s,t)$, the optimal function is found to be $\psi = \mathscr{R}^{-1}\mu$, where $\mathscr{R}f(\cdot) = \int \rho(\cdot,t)f(t)\mathrm{d}t$ and $\mu(t) = \mu_1(t) - \mu_0(t)$, corresponding to a misclassification error of $1 - \Phi(||\mathscr{R}^{-1/2}\mu||)$.

Now, depending on the interplay between the quantities $||\mathscr{R}^{-1}\mu||$ and $||\mathscr{R}^{-1/2}\mu||$ one can face different scenarios. Here we summarize the results:

- $||\mathscr{R}^{-1}\mu|| < \infty$ and $||\mathscr{R}^{-1/2}\mu|| < \infty$. Since the first quantity is bounded, the function $\psi(t)$ is well defined. The expression $\psi = \mathscr{R}^{-1}\mu$ permits to achieve the best possible error, which is some value greater than zero.

- $||\mathscr{R}^{-1}\mu|| = \infty$ and $||\mathscr{R}^{-1/2}\mu|| < \infty$. In this case the function ψ is not well defined. However, the best possible error can be achieved as the limit of classification errors computed along some sequence $\{\psi^{(k)}\}_{k\geq 1}$.

- $||\mathscr{R}^{-1}\mu|| = \infty$ and $||\mathscr{R}^{-1/2}\mu|| = \infty$. This is the most interesting case in the sense that the best possible error is zero. However, this error can be achieved again only as the limit of classification errors computed along some sequence $\{\psi^{(k)}\}_{k\geq 1}$.

In other words, even when the projecting function is not well defined, it is possible to achieve the minimum misclassification error considering the not convergent sequence $\{\psi^{(k)}\}_{k\geq 1}$. The nature of the problem is not a complication but rather an advantage in the sense that the more ill-posed the problem is the better optimal misclassification probability. Methods to define the sequence $\{\psi^{(k)}\}_{k\geq 1}$ will be clarified in the next chapter.

The *perfect classification* phenomenon is a unique feature of functional data. In some sense, the infinite dimensional nature of the these data translates in a greater amount of information with respect to the high dimensional (but not infinite) case.

1.3 Partially observed functional data

1.3.1 Problems and limits

As introduced in the previous sections, functional data are observed on a finite grid T_i. This grid can exhibit good features as, for example, some kind of regularity.

The most desirable case is when the grid is the same across subjects in the sample, i.e. $T_i = T$ for all i, as in figure 1.4. As anticipated in section 1.1.2, in this case the data can be easily stored in a matrix and computations can be done fastly.

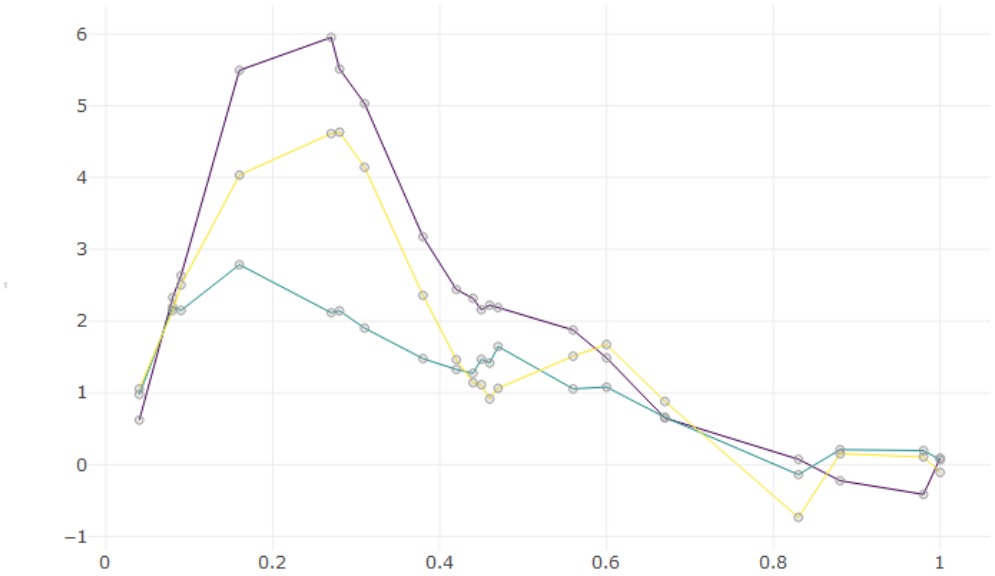

Figure 1.4: *Regular design for synthetic data. The three sample curves are observed on the same common grid.*

A second situation occurs when the grid is different from subject to subject but the starting and ending point of the grids are the same for all subjects, see figure 1.5. In this second case, for a fixed time point t_{ij} not all the curves are available. If we store the data in matrix we see some missing values induced by the irregularity of the grids. This creates several problems: estimation of the mean function and the covariance kernel as well as functional Principal Component Analysis are precluded due to missing data. However, these difficulties can be overcome using smoothing. More practically, one could smooth each curve individually and then re-evaluate each function on a common grid. This preprocessing procedure slight modifies the data but allow us to store the data in a matrix without missingness and then to use standard statistical procedures.

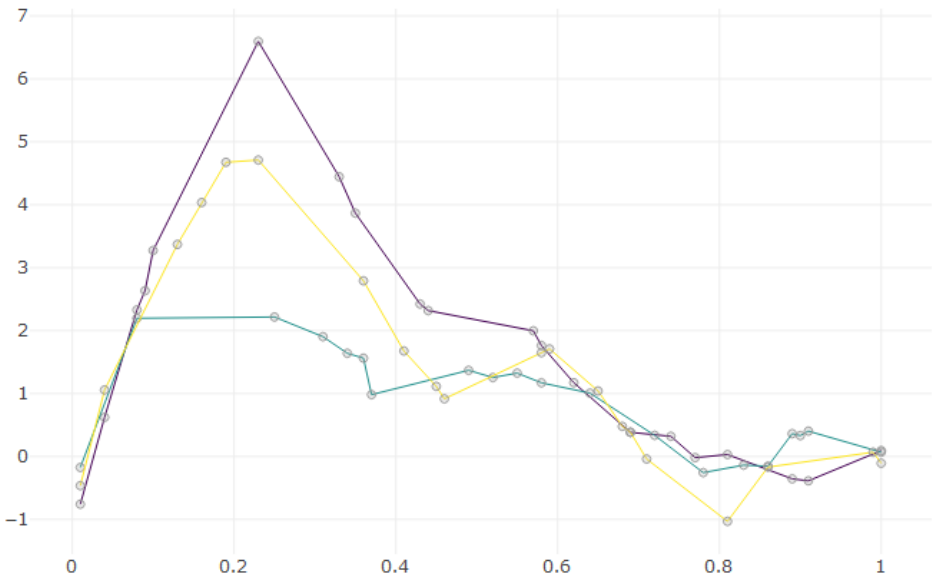

Figure 1.5: *Irregular design for synthetic data. The observation grids are subject-specific. Note that the starting and ending points of each curve are the same.*

Instead, consider the situation in which the starting and ending point of each grid T_i are different among subjects, like in figure 1.6. The reasons behind this fact can be various. For example, this feature is quite common in longitudinal data analysis, where subjects enter or drop the study at different times. In the dataset that we are going to analyze in the subsequent chapters, the curves start and end at different locations depending on where the medical scan has been centered. This last case is the most challenging: the problem cannot be solved using individual smoothing because it is not possible to interpolate the data where these are not observed at all. We refer to this case as *partially observed functional data* or, alternatively, as *functional fragments*. A survey of methods devoted to this kind of data are presented in the next subsection.

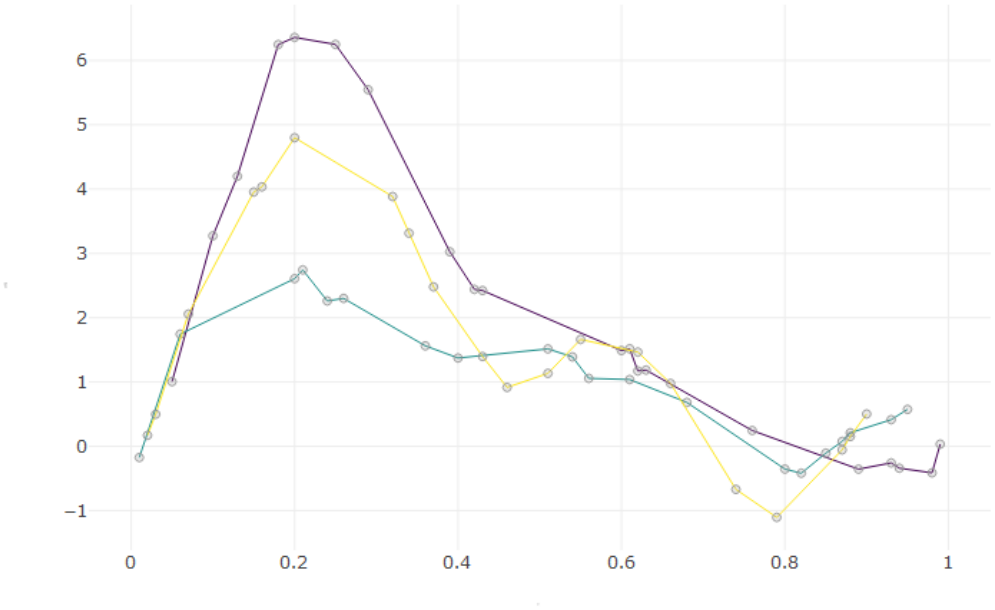

Figure 1.6: *Partially observed synthetic data. The observation grids and the starting and ending points are subject-specific.*

1.3.2 Possible approaches to functional fragments

Since the beginning of 2000s researchers focused the attention on partially observed functional data. The presence of the missingness in this kind of data can be faced at different levels, depending on the goal of the analysis. At the most general level one possibility is to impute the unobserved parts of the curves in the sample in order to have a complete dataset. Once the imputation is done, standard statistical procedures can be applied. Several techniques have been developed under this framework: Goldberg et al. (2014) propose a method to predict the continuation of a curve based on the Best Linear Unbiased Predictor (BLUP) and a B-spline representation; the works by Delaigle and Hall (2013) and Delaigle and Hall (2016) introduce different approaches to the problem, the first one being a nonparametric curve extension method strictly related to the task of classification while the second one a markov chain – based technique to forecast the unobserved parts of the curves; Kraus (2015) face the problem using a linear approximation of the conditional expectation between the unobserved and observed part of the curve. This

last approach is further developed in an upcoming work by Alois Kneip and Dominik Liebl that can be found in arXiv (`https://arxiv.org/abs/1710.10099`). The issue of the violation of the missing-completely-at-random assumption in the imputation procedure is discussed in Liebl and Rameseder (2018).

A second level of imputation consist in reconstructing the covariance kernel and not the whole trajectories. This approach is justified by the fact that several models and methods only require this kind of quantities, i.e. averages across subjects, and do not need individual values. In this framework the most remarkable work is Yao et al. (2005). The authors propose a new algorithm called PACE (Principal Analysis via Conditional Expectation) that reconstructs the covariance surface using local linear smoothers. This method has been successful and has been applied several times over the years: for example Liebl (2013) apply PACE in the context of prediction of electricity prices and Gromenko et al. (2017) extend the methodology to the case of spatially correlated data. Under the same framework, also Kraus (2015) develop its own method: it basically uses only available curves on each timepoint couple (t_j, t_k). These two methodologies will be discussed in the subsequent chapters. It is worth mentioning the upcoming paper by Marie-Hélène Descary and Victor Panaretos that can be found in arXiv (`https://arxiv.org/abs/1708.02491`): the authors use matrix completion techniques to define a nonparametric consistent estimator and reconstruct the covariance surface.

When the aim is functional Principal Component Analysis, the reconstruction of the covariance function is not really mandatory. In fact, there are few methods that extracts the eigenfunctions and the scores without the computation of the covariance kernel. The literature on longitudinal data analysis was the inspiration for the works by James et al. (2000) and James and Hastie (2001). Making use of a linear mixed model together with a B-spline basis representation, the authors are able to construct a likelihood-based method imputing both the eigenfunctions and the scores in the first case and employing a discriminant analysis for fragments in the second case. Again, the works by Yao et al. (2005) and Kraus (2015) are able to predict the scores: the first one using conditional expectations under the assumption of gaussianity while the second one using best linear approximation of the conditional expectation. Lastly, the original method by Huang et al. (2008),

developed in the context of fully observed trajectories, can be adapted to the case of partially observed data, as explained in Lila et al. (2016).

Chapter 2

PCA–based discrimination of partially observed functional data, with an application to Aneurisk65 dataset

2.1 Introduction

The AneuRisk data (`https://statistics.mox.polimi.it/aneurisk/`), displayed in figure 2.1, consist in the profiles of radius (left) and curvature (right) of the internal carotid artery of 65 subjects (see, e.g., Sangalli et al., 2009a, 2014b). The data originate from the reconstruction of three-dimensional angiographic images, taken on subjects suspected to be affected by cerebral aneurysms. The domain where each datum is observed varies across subjects, with longer or shorter portions of the internal carotid artery being observed, depending on where the medical scan has been centered. As highlighted in the figure, there is one portion of the domain where all the data are observed; this corresponds to the (approximately) 3 cm closer to the terminal bifurcation of the artery, that is a point of specific clinical interest; on the other hand, for most subjects, longer portions of the artery are observed, up to more than 10 cm. This incomplete data setting, where there is one portion of the domain where all data are observed, but individual observations are progressively lost when moving from this portion of the domain towards the full domain, is common in functional data coming from medical imaging and from

biological studies in general.

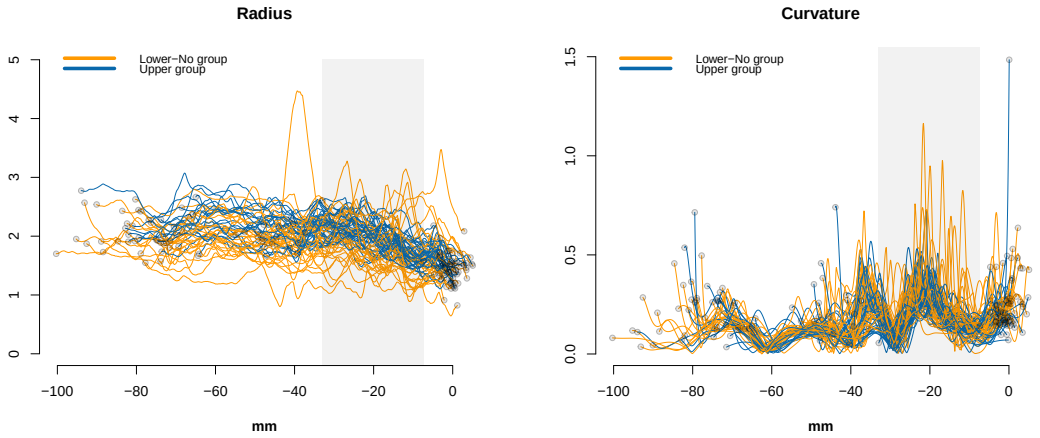

Figure 2.1: *The Aneurisk65 dataset. Registered radius (left) and curvature (right) of the internal carotid arteries of 65 subjects. The portion of the domain where the data are observed for all subjects is highlighted in light–gray. The circles indicates the starting and ending points for each datum. Two different colors are used for subjects in the Upper group (blue) and subjects in the Lower–No group (orange).*

The analysis of AneuRisk65 data is relevant for the study of the pathology of cerebral aneurysms; in particular, it is relevant to investigate whether the morphology of the internal carotid artery influences aneurysms pathogenesis. The data can be divided into two groups, displayed in orange and blue in the figure, depending on the presence and location of the cerebral aneurysms. In particular, 33 subjects have an aneurysm at or after the terminal bifurcation of the internal carotid artery (Upper group) while the remaining 32 subjects, either have an aneurysm along the internal carotid artery, before the terminal bifurcation, or were found no apparent aneurysm during the angiography (these 32 subjects compose the Lower–No group). Sangalli et al. (2009a) present a discriminant analysis between these two groups, based on the scores of the principal components of the radius and curvature profiles; in the latter work, the principal components are computed restricting the attention to the portion of the domain common across subjects. It is however natural to wonder whether these discrimination results may be improved by considering also portions of the domain where not all data are observed.

Unfortunately, most of the nowadays very extensive literature on functional data analysis focuses on the case where all functional data are observed over a common domain, and the vast majority of functional data analysis techniques so far developed is not able to handle this incomplete data framework. The development of methods for partially observed functional data has thus recently started attracting an increasing interest. Classification of functional fragments is discussed in James and Hastie (2001) where an extension of the linear discriminant analysis to the incomplete data framework is proposed. An alternative discrimination technique, based on curves extension, is presented by Delaigle and Hall (2013) and further developed in Delaigle and Hall (2016). Methods for functional Principal Component Analysis (fPCA) of incomplete functional data are described for instance in James et al. (2000), Yao et al. (2005) and in Kraus (2015). Di et al. (2014) extend the technique by Yao et al. (2005) to a multilevel setting, while Liu et al. (2017) employ it to handle spatio–temporal data with gaps. Other works consider partially observed functional data in different applied contexts: Liebl (2013) develops a functional factor model for electricity spot prices, Goldberg et al. (2014) focus on curve forecasting for call center data, and Gromenko et al. (2017) propose a functional regression model for physical data.

Here, in particular, we shall focus on discrimination based on fPCA scores and consider the case where the incomplete functional data share one common portion of the domain, likewise AneuRisk65 data. The natural usage of techniques for incomplete functional data consists in applying the technique to the whole domain where at least one functional datum is observed. However, this may not be the optimal choice, especially for classification purposes. We will specifically show that, when considering discrimination based on fPCA scores, enlarging the analysis to the whole domain, as well as restricting it to the common domain where all data are observed, may not lead to the best classification results. As illustrated via a simulation study and an application to AneuRisk65 data, the optimal choice often lies between these two extremes. We here suggest to explore different extensions of the domain, ranging from the common domain to the full domain, and select the one that provides the best discrimination result under cross–validation.

Section 2.2 reviews the techniques for fPCA of incomplete functional data proposed by James et al. (2000), Yao et al. (2005) and Kraus (2015). The same

section also generalizes to the incomplete data setting the regularized fPCA technique originally proposed by Huang et al. (2008) in the completely observed data scenario. Section 2.3 describes the domain extension approach for fPCA–based discrimination. Section 2.4 illustrates this idea in a simulation study while Section 2.5 shows the application to AneuRisk65 data. Finally, Section 2.6 draws some concluding remarks and outlines future directions of possible research.

2.2 fPCA of partially observed functional data

Assume that N functional data $x_1(t), \ldots, x_N(t)$ are generated from some real-valued random process $Z(t)$, with mean $\mu(t)$ and covariance kernel $\rho(s,t)$, and that only a discrete and noisy version of each datum is available, i.e., $x_{ij} = x_i(t_{ij}) = z_i(t_{ij}) + \varepsilon_{ij}$ for $i \in \{1, \ldots, N\}, j \in \{1, \ldots, m_i\}$, where ε_{ij} are measurement errors, with zero mean and finite variance. Consider in particular the case where the observation grids $\{t_{i1}, \ldots, t_{im_i}\}$, with $t_{i1} < \cdots < t_{im_i}$, may differ over the various statistical units, $i = 1, \ldots, n$, and that the domains where they insist, $T_i = [t_{i1}, t_{im_i}] \subset \mathbb{R}$, may as well be different. Standard fPCA assumes the representation

$$z_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\varphi_k(t), \quad i \in \{1, \ldots, N\}, \tag{2.1}$$

where $\mu(t)$ is the mean function, $\varphi_k(t)$ is the $k^{\texttt{th}}$ eigenfunction of the covariance kernel $\rho(s,t)$ and ξ_{ik} is the corresponding score for the $i^{\texttt{th}}$ observation. In practice, only the first K elements of the series are considered. In particular, when the data are completely observed over a common domain $T = [t_1, t_m]$ and on a common grid $\{t_1, \ldots, t_m\}$, the same for all statistical units, the first K principal components can be estimated performing the eigendecomposition of the empirical covariance matrix; the corresponding scores, theoretically defined as $\xi_{ik} = \int Z_i(t)\varphi_k(t)\,\mathrm{d}t$, for each i and k, can be computed by discretizing the integral. When the observation grid differs across the statistical units, but the domain is common to all units, i.e. $T_i = T$, one possibility is to smooth separately each functional datum, and then evaluate each function on a new regular grid, common to all statistical units. Unfortunately, when the data are only partially observed, or observed over different domains T_i, the individual smoothing is not useful for inferring the values of the

functions where these are not observed; hence, it is not possible to compute the principal components and associated scores as described above. In this situation, few methodologies try to estimate the scores and the eigenfunctions considering different reformulation of the estimation problem.

James et al. (2000)

Mixed effect models are widely used to handle missing data in longitudinal data analysis. Borrowing from these approaches, James et al. (2000) propose a mixed effect model where the principal component scores are treated as random effects and the mean and principal components are represented via a spline basis. Denote by $\eta(t)$ a spline basis with dimension q. The mean and principal components are then represented as $\mu(t) = \eta(t)^{\mathtt{T}}\mathbf{c}_\mu$ and $\mathbf{f}(t)^{\mathtt{T}} = \eta(t)^{\mathtt{T}}\mathbf{C}$, where $\mathbf{c}_\mu$ and $\mathbf{C}$ are, respectively, a q–dimensional vector of spline coefficients and a $(q \times K)$ matrix of spline coefficients. From equation (2.1), this leads to the model

$$x_i(t) = \eta(t)^{\mathtt{T}}\mathbf{c}_\mu + \eta(t)^{\mathtt{T}}\mathbf{C}\,\mathbf{u}_i + \zeta_i(t),$$

where the $\mathbf{u}_i$s are assumed to have zero mean and a common variance $\widehat{\rho}_u$, and the $\zeta_i(t)$s are assumed to have zero mean and a constant variance function σ^2. To ensure identifiability of $\mathbf{C}$ and $\widehat{\rho}_u$ the authors restrict the covariance matrix of the $\mathbf{u}_i$s to be diagonal. The fitting procedure is based on maximum likelihood estimation and makes use of the EM algorithm. Once the estimates of the principal components are obtained, the estimates of the scores can be computed through best linear unbiased prediction. The number of basis functions acts as a smoothing parameter that must be carefully selected.

Yao et al. (2005)

Yao et al. (2005) develop an algorithm called PACE (Principal Analysis via Conditional Expectation) that estimates the principal component scores using conditional means. They first estimate the mean $\mu(t)$ and the covariance function $\rho(s,t)$ via local linear smoothing of the raw mean vector and covariance matrix

obtained from pooled data. An important choice in this context is the selection of the bandwidths h_μ and h_ρ for the two kernel smoothers. Once the estimates $\widehat{\mu}(t)$ and $\widehat{\rho}(s,t)$ of the mean and covariance functions are available, the estimates $\{\widehat{\varphi}_1, \ldots, \widehat{\varphi}_K\}$ of the first K principal components are determined solving the usual discretized eigenvalue–eigenfunction problem, with the associated estimated eigenvalues $\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K\}$. The best prediction for the score vector $\boldsymbol{\xi}_i$, associated with the $i^{\texttt{th}}$ observation $\boldsymbol{x}_i = \big(x_{i1}, \ldots, x_{im_i}\big)^{\texttt{T}}$, is the conditional expectation given $(\boldsymbol{Z}_i = \boldsymbol{x}_i)$. Under Gaussian assumptions for the measurement errors ε_{ij} and for the scores themselves, this can be shown to be

$$\widehat{\xi}_{ik} = \widehat{\mathbb{E}}[\xi_{ik} |\, \boldsymbol{Z}_i = \boldsymbol{x}_i] = \widehat{\lambda}_k\, \widehat{\boldsymbol{\varphi}}_{ik}^{\texttt{T}}\, \widehat{\rho}_i^{-1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_i),$$

where $\widehat{\boldsymbol{\varphi}}_{ik} = \big(\widehat{\varphi}_k(t_{i1}), \ldots, \widehat{\varphi}_k(t_{im_i})\big)^{\texttt{T}}$, $\big[\widehat{\rho}_i\big]_{j\ell} = \widehat{\rho}(s_{ij}, t_{i\ell})$, and $\widehat{\boldsymbol{\mu}}_i = \big(\widehat{\mu}(t_{i1}), \ldots, \widehat{\mu}(t_{im_i})\big)^{\texttt{T}}$ are computed on each individual grid T_i.

Huang et al. (2008)

In the case of completely observed functional data, Huang et al. (2008) propose a regularized version of fPCA, that can be be easily generalized to partially observed data, as noted in Lila et al. (2016). This approach relies on a different characterization of the principal components, the so–called best K bases approximation property. Namely, the first K principal components enable the best reconstruction of the signals, in an L^2 sense, among all orthonormal bases of dimension K:

$$\big\{\varphi_k\big\}_{k=1}^{K} = \underset{\left\{\{\psi_k\}_{k=1}^{K} : \int \psi_s, \psi_l = \delta_{sl}\right\}}{\operatorname{argmin}} \mathbb{E}\left[\int \left\{Z - \mu - \sum_{k=1}^{K} \left(\int X \psi_k\right) \psi_k\right\}^2\right].$$

Considering only one principal component, the empirical version of the expectation above, for partially observed functional data, is given by $\sum_{i=1}^{n} \sum_{j=1}^{m_i} \big(x_{ij} - \xi_i \varphi(t_{ij})\big)^2$. Since the minimization of this quantity involves raw data, a roughness penalty on φ is introduced to ensure smoothness of the resulting principal component. In particular, the first principal component and the associated score vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{\texttt{T}}$ are estimated solving the following minimization problem:

$$\underset{\boldsymbol{\xi}, \varphi}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \big\{x_{ij} - \xi_i \varphi(t_{ij})\big\}^2 + \gamma\, \boldsymbol{\xi}^{\texttt{T}} \boldsymbol{\xi} \int \big\{\varphi(t)\big\}^2 \,\mathrm{d}t.$$

The smoothing parameter $\gamma > 0$ controls the regularity of the estimated principal component $\varphi(t)$. The term $\boldsymbol{\xi}^{\mathtt{T}}\boldsymbol{\xi}$ is included to obtain desirable invariance properties (see Huang et al., 2008, for details). Subsequent principal components and the associated score vectors are estimated sequentially solving the same minimization problem, once the contribution to the data of the previously estimated principal components is removed.

Kraus (2015)

Kraus (2015) shows another way to deal with the problem of fPCA in the case of partially observed functional data. The starting point is to estimate the mean function $\mu(t)$ using, for each t, only the available curves at the specific time t, and to estimate $\rho(s,t)$ using all complete pairs of functional values at s and t. It is shown that under technical conditions concerning the information provided by the observation grids, these estimators are consistent. The eigenfunctions of the covariance operator can be estimated performing spectral analysis of the complete pairs sample covariance. The missing part of each score is predicted via best linear approximation of its conditional expectation. Using the Riesz representation theorem, the optimization problem can be written as

$$\min_{a_{ik} \in L^2(T_i)} \mathbb{E}\left[\left(\xi_{ik,mis} - \int_{T_i} a_{ik} x_i\right)^2\right]$$

where $\xi_{ik,mis}$ is the missing part of the $k^{\mathtt{th}}$ score for the $i^{\mathtt{th}}$ unit, a_{ik} is an element of $L^2(T_i)$ and x_i is the observed curve. This leads to a linear inverse problem that is regularized, thus involving also in this case the choice of a regularization parameter. Note that this methodology assumes that the data are observed without noise. Moreover, the technique can only deal with data observed over grids that, apart for the starting and ending points, are common across statistical units. This does not create problems in the application to Aneurisk65 data, as these data are preprocessed and evaluated on a common regular grid (see Sangalli et al., 2014b, for details on the preprocessing). In general, a pre-smoothing of each functional data and the re-evaluation on a common grid may be necessary before the technique by Kraus (2015) can be implemented.

2.3 PCA–based discrimination of partially observed functional data

The four methods for fPCA of incomplete functional data, briefly reviewed in Section 2.2, are based on different estimation problems and is not clear in advance which one is preferable and in which situation. The first two models rely on parametric assumptions, while the third and the fourth do not. The first three methods involve smoothing, but in different ways: James et al. (2000) use a B–spline basis to represent the mean and principal components, Yao et al. (2005) use a kernel smoothing for the mean and the covariance function, and Huang et al. (2008) smooth the eigenfunctions using a roughness penalty approach. For all the methods, one or more tuning parameters must be selected in some optimal way: the number of B–spline basis in James et al. (2000), the two kernel bandwidths in Yao et al. (2005), the smoothing parameter γ in Huang et al. (2008), and the regularization parameter in Kraus (2015).

In the following sections, we use these methods to perform discrimination of partially observed functional data, where the discrimination is based on the scores of the first K principal components. One natural approach in this sense would be to carry out the analysis over the full domain. However, we show that working on the largest possible domain may not be the optimal choice for classification purposes. On one hand, this may result in imprecise estimates of the principal components, especially of high order, where many data are missing. On the other hand, when the target is classification, considering the total domain may not be useful, when most of the between–group variability is located within the common domain or close to it, or when the missingness is so important that is difficult to distinguish between–group and within–group variability.

We here instead suggest to explore different portions of the domain, moving from the common domain and progressively enlarging towards the full domain. More specifically, we divide the domain where the data are partially observed in L portions, and we thus consider a collection of progressively larger domains $\mathcal{I}_\ell$ for $\ell \in \{0, \ldots, L\}$, with $\mathcal{I}_{\ell-1} \subset \mathcal{I}_\ell$, where $\mathcal{I}_{\texttt{min}}$ is the common domain and $\mathcal{I}_{\texttt{max}}$ is the full domain. Figure 2.2 shows such domains extensions for the AneuRisk65 data. The principal components and their associated scores are then computed over each

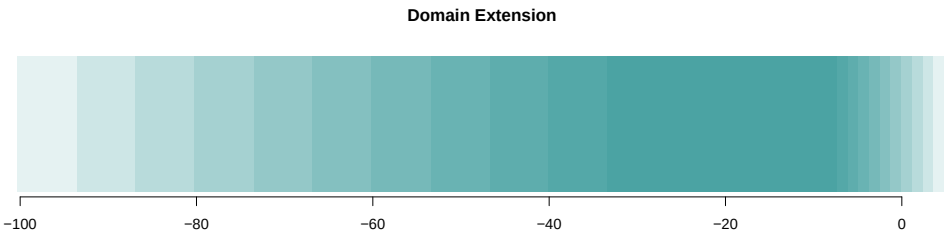

Figure 2.2: *Visual illustration of the domain extensions for AneuRisk65 data. Moving from the portion of the domain where we have observations for all statistical units, i.e., the common domain, here highlighted by the darkest color, we progressively enlarge the domain by constant steps, until we reach the full domain, where at least one statistical unit is observed, here indicated by the lightest color. The various domain extensions are denoted by progressively lighter shades of color.*

domain extension $\mathcal{I}_\ell$, and used for the classification. In particular, in the following sections we consider quadratic discriminant analysis (see, e.g., Izenman, 2009) on the scores of the first K principal components. The optimal number of principal components and the optimal domain extension $\mathcal{I}_\ell$ are selected via cross–validation. For simplicity, the domains $\mathcal{I}_\ell$ are defined by constant enlargements from the common domain to the full domain. Moreover, for illustrative purposes, we here carry out an exhaustive search from $\mathcal{I}_{\texttt{min}}$ to $\mathcal{I}_{\texttt{max}}$. Of course, the enlargement step as well as the search could be optimized, if necessary, to decrease the computational cost.

2.4 Simulations

To illustrate the domain extension approach we carry out a simple simulation study. We generate a set of $N = 100$ functional data over the interval $\mathcal{I}_{\texttt{max}} = [0, 1]$. We then completely retain the data generated over the interval $\mathcal{I}_{\texttt{min}} = [1/3, 2/3]$, while we censor them over the intervals $\mathcal{I}_{\texttt{left}} = [0, 1/3]$ and $\mathcal{I}_{\texttt{right}} = [2/3, 1]$, by sampling the starting point of each functional datum uniformly over $\mathcal{I}_{\texttt{left}}$, and its ending point uniformly over $\mathcal{I}_{\texttt{right}}$. For four statistical units the starting or ending observation points are not sampled but fixed, so that we have one functional datum with starting point in 0, another with starting point in 1/3,

one functional datum with ending point in 2/3 and another with ending point in 1; this ensures that the full domain is $\mathcal{I}_{\texttt{max}} = [0, 1]$ and the common domain is $\mathcal{I}_{\texttt{min}} = [1/3, 2/3]$. The data are generated from a cubic B–splines basis with 16 internal knots, corresponding to a total of 20 bases. The position of the spline knots is displayed in Figure 2.3 by small vertical markers along the x-axis. We generate two groups of functional data, $g \in \{1, 2\}$, composed by 50 curves each, by sampling at each simulation repetition the spline coefficients $\{c_{1,g}, c_{2,g}, \ldots, c_{20,g}\}$ for the two groups from normal distributions with group–specific means, $c_{s,g} \sim N(\mu_{s,g}, \sigma^2)$, for $s \in \{1, \ldots, 20\}$. The means of the first group, $\{\mu_{1,1}, \mu_{2,1}, \ldots, \mu_{20,1}\}$, are set equal to $\{0, 0, 0, 0, 1, 2, 1, 0, -1, 2, 2, -1, 0, 0.5, 1, 0.5, 0, 0, 0, 0\}$; the means of the second group are set to $\{\mu_{1,2}, \mu_{2,2}, \ldots, \mu_{20,2}\} = \{\mu_{20,1}, \mu_{19,1}, \ldots, \mu_{1,1}\}$, thus taking the same values as the first group, but in reverse order. The difference in the means of the spline coefficients constitutes the only structural difference between the two groups. The variance σ^2 of the spline coefficients is the same in both groups and across different coefficients and is set to $\sigma^2 = 0.6$. The N generated curves are evaluated on a regular grid of $p = 150$ over $[0, 1]$ and contaminated by additive, uncorrelated, Gaussian noise, with mean zero and constant variance $\sigma^2_\varepsilon = 0.1$. This simulation is repeated 50 times. Different simulation settings are considered in the appendix, changing the amount of noise, the variance of the spline coefficients, the mean values of the coefficients. Figure 2.3 shows the data sampled in the first simulation repetition. The discrimination between the two groups of data is present both within and outside the common domain, with an important part of the discrimination lying outside the common domain.

We thus apply the three methodologies for fPCA of partially observed functional data reviewed in Section 2.2, over 10 domain extensions, ranging from $\mathcal{I}_{\texttt{min}}$ to $\mathcal{I}_{\texttt{max}}$, with constant enlargement steps. The analysis is performed in the `R` environment (R Core Team (2016)). The tuning parameters of each methodology are selected at each simulation replicate by cross–validation. This is carried out separately over each domain extension; the selected tuning parameters can thus differ for different domain extensions. The selection of the optimal number of spline bases in James et al. (2000), implemented in the `R` package `fpca` (Peng and Paul, 2011), is carried out optimizing an approximate cross–validation score. For Yao et al. (2005), implemented in the package `fdapace` (Dai et al., 2017a), the

Figure 2.3: *Simulation study 1. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

optimal bandwidths for the two kernel smoothers are chosen minimizing the leave–one–curve cross validation. For the method based on the extension of Huang et al. (2008) to partially observed data, we implemented a 5–fold cross–validation. The regularization parameter in Kraus (2015), implemented through routines published by the author [1], is selected via generalized cross validation.

A quadratic discriminant analysis on the scores of the first K principal components, with $K \leqslant 5$, is then carried out. In particular, for each replication and each domain extension, we select the optimal number of principal components scores to be considered for the discrimination via leave–one–out cross–validation, but minimizing in this case the misclassification error. We also apply standard PCA to the fully observed (non–censored) data; the associated misclassification error indicates the best possible classification results achievable in this simulation setting, based on discrimination of the principal component scores, for fully observed data. The top panel of Figure 2.4 displays the boxplots of the leave-one-out misclassification error, for the various techniques, for various domain extensions. The leave-one-out misclassification error that could be attainable if the uncensored

[1] available at `http://dx.doi.org/10.1111/rssb.12087`

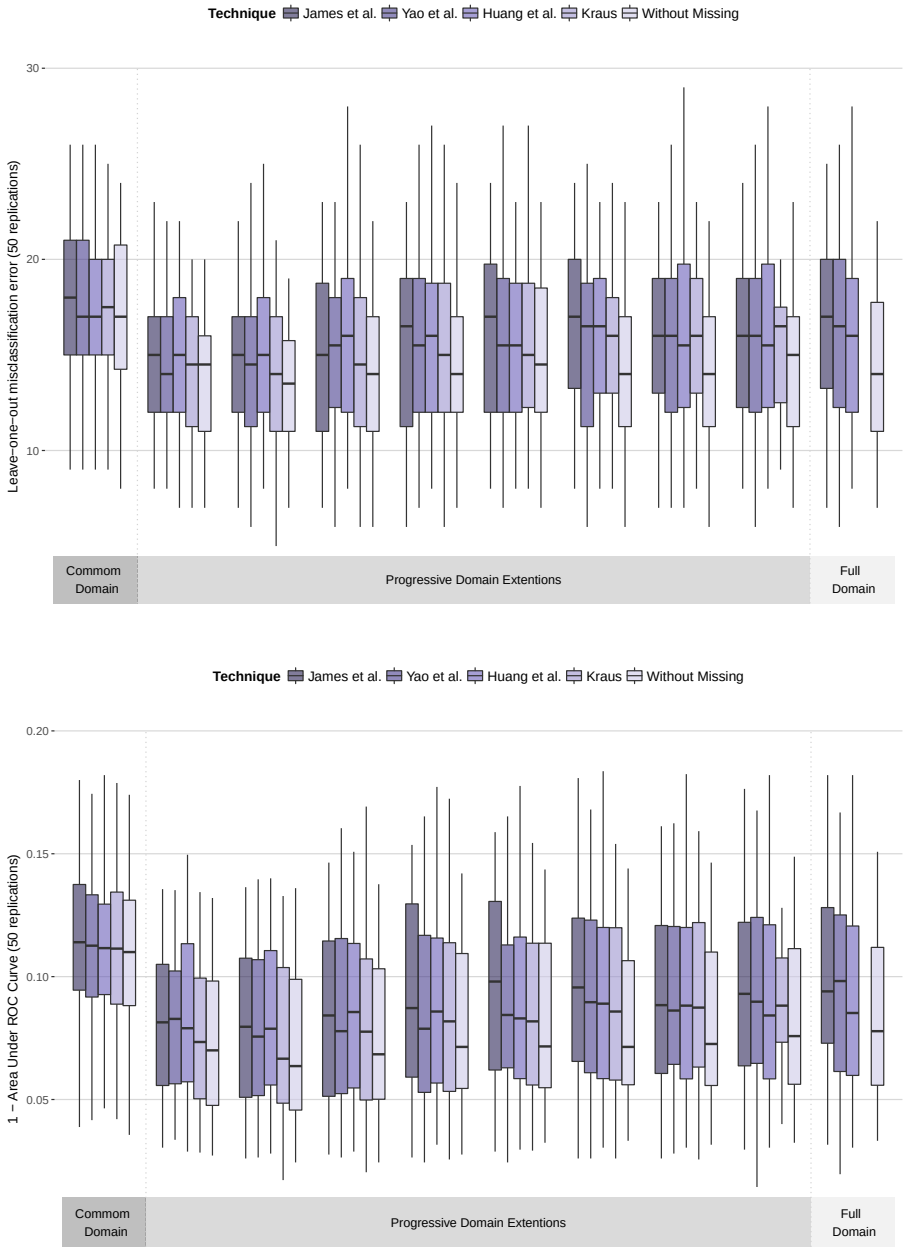

Figure 2.4: *Simulation study 1. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

data were available is as well displayed. Note that in the full domain the method by Kraus (2015) is not employable because there are no curves observed jointly at time 0 and 1. For all methods, the misclassification error decreases when we start extending the domain with respect to the common domain, but then progressively increases as we approach the full domain. None of the methods outperforms the other. As an additional measure of the quality of the discrimination we also compute the area under the ROC Curve (see Izenman (2009)); this quantity is bounded between 0 and 1, with the value 1 being attained for perfect classification. In the bottom panel of Figure 2.4 we show the boxplots of the index (1 - area under the ROC curve), whose minima correspond to the best discrimination. A visual inspection of these boxplots confirms what already commented on the base of the leave-one-out miscassification error: extending the domain with respect to the common domain improves the discrimination between the two groups; on the other hand, larger domain extensions, and in particular the full domain, do not lead to the best discrimination results.

2.5 Application to AneuRisk65 data

The AneuRisk project (`https://statistics.mox.polimi.it/aneurisk/`) is an interdisciplinary project that involved statisticians and numerical analysts from Politecnico di Milano (Milano, Italy) and Emory University (Atlanta, USA), bioengineers and computer scientists from Istituto Mario Negri (Bergamo, Italy), and medical doctors from Niguarda Hospital and Maggiore Policlinico Hospital (Milano, Italy), with the aim of investigating cerebral aneurysms pathology. This is a very common pathology, totally asymptomatic in the vast majority of cases. Rupture of a cerebral aneurysm is a rare event (affecting one in ten thousand people every year), but unfortunately has associated very high mortality. The origin of the pathology is still largely unknown. One conjecture, investigated by the AneuRisk project, is that aneurysm's pathogenesis may be influenced by the morphology of the hosting vessels, and in particular by the morphology of the internal carotid artery, through the effect that the morphology of the vessel has on the blood fluid–dynamics. The two geometrical quantities that mostly determine

the haemodynamics are the radius and the curvature of the vessel. For this reason, the first studies carried out within the AneuRisk project focused on these two features. Figure 2.1 shows the profiles of radius (left) and curvature (right) of the internal carotid artery of 65 subjects, pre-processed and registered as described in Sangalli et al. (2009a, 2014b,a). As outlined in section 2.1, the data are divided in 2 groups depending on the presence and location of the aneurysm. Sangalli et al. (2009a) carry out a discriminant analysis between these two groups, based on the scores of the principal components of the radius and curvature profiles, computing the principal components by standard fPCA on the portion of the domain common across subjects (the approximately 3cm closer to the terminal bifurcation of the internal carotid artery, as highlighted in Figure 2.1). The resulting leave–one–out misclassification error amounts to 15 subjects.

Here we test the four methodologies described in the previous sections over various domains extensions; see Figure 2.2 for the considered domain extensions. Likewise in Sangalli et al. (2009a), we consider up to 4 principal components. As for the simulation, the optimal number of principal components is selected for each method and each domain extension via leave-one-out cross validation, minimizing the misclassification error. Figure 2.5 shows the classification results.

For all considered methods, the domain where the best discrimination is achieved lies between the common domain and the total domain. In this particular application, the approach based on the extension of Huang et al. (2008) to partially observed functional data does the best job, reaching a leave–one–out misclassification error of 9 subjects. Huang et al. (2008) returns the best results also when considering the index based on the area under the ROC curve. Looking at the leave-one-out missclassification error, the best domain extension for this method turns out to be optimal also for the other techniques considered. On the common domain, all methods perform similarly to standard fPCA, with 14, 15 or 16 misclassified units, depending on the method. As highlighted by this figure, the application of the methodologies for partially observed data on the full domain does not lead to any improvement in the discrimination; for discrimination based on James et al. (2000) and Yao et al. (2005), the misclassification error is in fact higher on the full domain than on the common domain, and the index based on the area under the ROC curve is as well worse on the full domain than on the

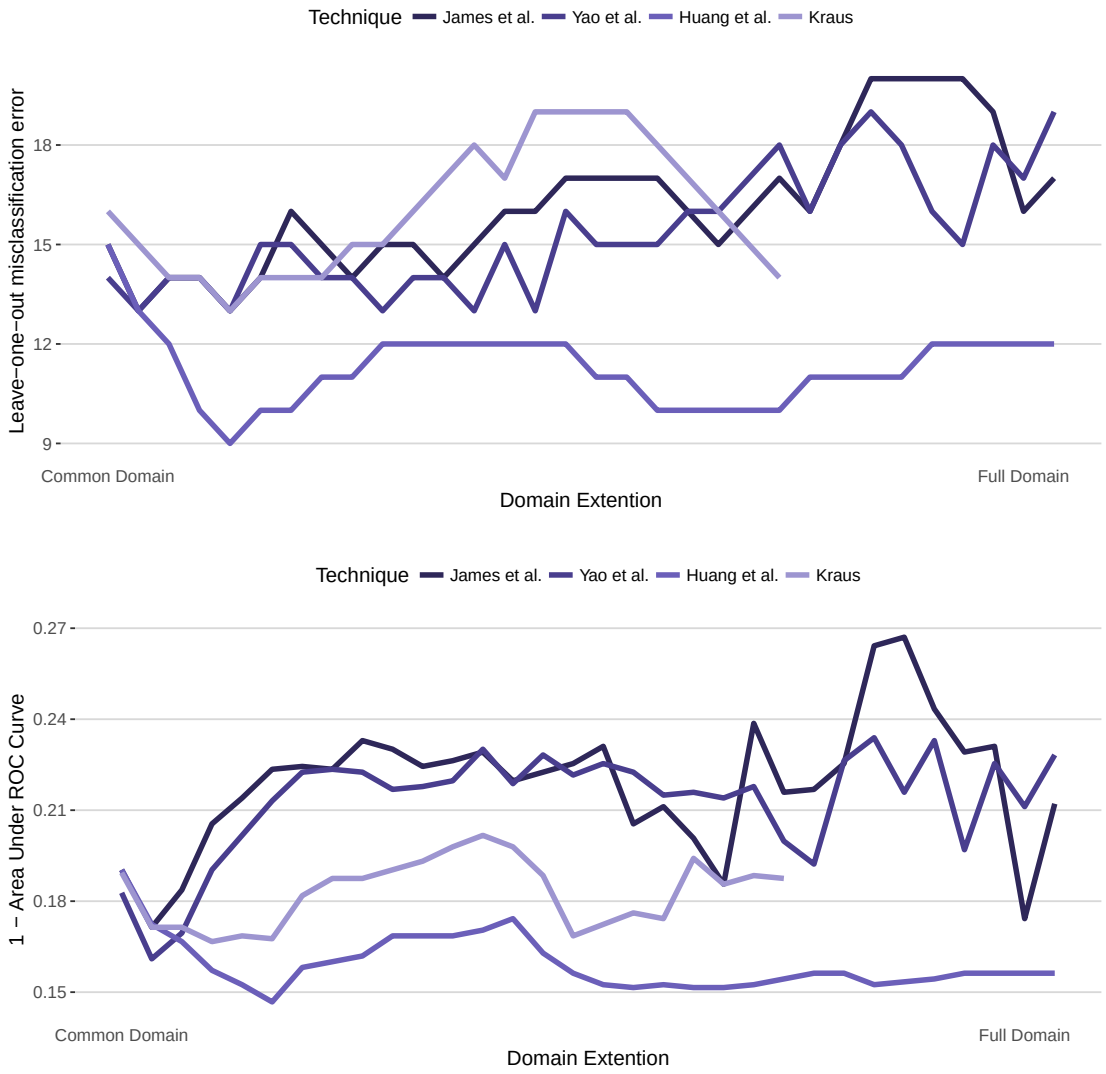

Figure 2.5: *AneuRisk65 data. Top: leave–one–out misclassification error for various domain extensions. Bottom:* 1− *Area Under ROC Curve for various domain extensions.*

common domain. So, ignoring the domain extension technique would lead to the incorrect conclusion that there is no advantage in including the part of the domain where the data is only partially observed.

The estimated principal components over the best domain extension in terms of misclassification error are displayed in Figure 2.6. The estimates of the principal components returned by the four methods are very similar. The second component

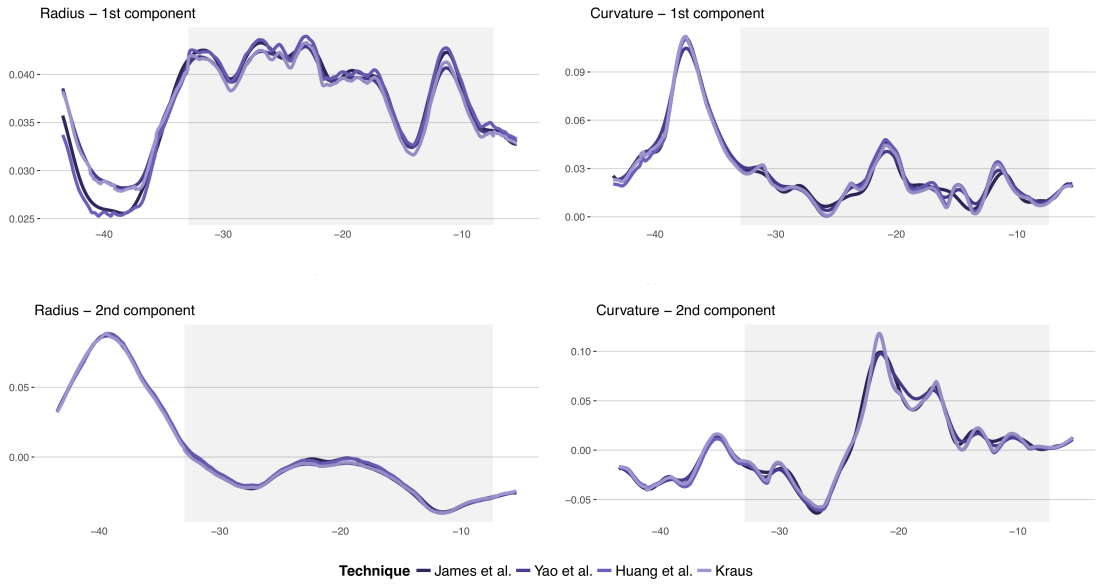

Figure 2.6: *AneuRisk65 data. Estimates of the principal components provided by the various considered methods on the optimal domain extension. The portion of the domain where the data are observed for all subjects is highlighted in light–gray.*

for the radius and the first for the curvature have important peaks outside of the common domain (at about −40mm and −38mm, respectively). An important part of the discrimination between the two groups lies here, and for this reason a better classification is possible only when considering the domain extension.

2.6 Discussion

As highlighted by the simulation study and the application to AneuRisk65 data, when performing supervised classification of partially observed functional data, considering the full domain where the data are observed may not be optimal. In this illustrated review of PCA–based discrimination of partially observed data, we explored a simple strategy of searching over domain extensions, moving from the common domain where all the data are observed to the full domain where at least some datum is available. An interesting line for future investigation goes towards a more complex and complete search for such optimal domain, where the search is not restricted to progressive extensions of the common domain. In

the context of fully observed functional data, a similar idea is introduced by Pini and Vantini (2016), futher developed by Pini and Vantini (2017) for supervised profile monitoring and explored in Floriello and Vitelli (2017) for unsupervised clustering. The domain–selection idea we are here considering differs instead from the approaches explored in Ferraty et al. (2010) and Delaigle et al. (2012), where the search focuses on specific pointwise locations where discrimination between two groups of functional data is optimized.

Chapter 3

Classification of functional fragments by regularized linear classifiers with domain selection

3.1 Introduction

We consider classification of a functional observation into one of two groups. Classification of functional data is a rich, long-standing topic comprehensively overviewed in Baíllo et al. (2011b). As pointed out in Chapter 1, Delaigle and Hall (2012a) show that depending on the relative geometric position of the difference of the group means, representing the signal, and covariance operator, summarizing the structure of the noise, certain classifiers can have zero misclassification probability. This remarkable phenomenon, called perfect classification, is a special property of the infinite-dimensional setting and cannot occur in the multivariate context, unless in degenerate cases. Delaigle and Hall (2012a) show that a particularly simple class of linear classifiers, based on a carefully chosen one-dimensional projection of the function to classify, can achieve this optimal error rate either exactly or in the limit along a sequence of approximations. Berrendero et al. (2017) further elucidate the perfect classification phenomenon from the point of view of the Feldman–Hájek dichotomy between mutual singularity or absolute continuity between two Gaussian measures on abstract spaces.

Motivated by these findings we reformulate the problem of determining the best classifier as a quadratic optimization problem on a function space, or, equiv-

alently, as a linear inverse problem. These problems are ill-posed which, unlike in most inverse problems, is not a complication but rather an advantage in the sense that the more ill-posed the problem is the better optimal misclassification probability. We use regularization techniques, such as the numerical method of conjugate gradients with early stopping and ridge regularization, to solve the optimization problem which leads to a class of regularized linear classifiers. The optimal misclassification rate is the limit along the regularization path of solutions which themselves may not converge.

We study the empirical version of the problem, where the objective function in the constrained minimization must be estimated from finite training data. Here our contribution is in two important aspects. First, we show that it is possible to construct an empirical regularization path towards the possibly non-existent unconstrained solution so that the classification error converges to its best, possibly zero, value. We do this specifically for three methods, namely conjugate gradient, principal component and ridge classification, in a truly infinite dimensional manner in the sense that the convergence takes place along a path with decreasing amount of regularization and holds without particular restrictions on the mean difference between classes. Second, all our methodology and theory is developed in the setting of partially observed functional data, where trajectories are observed only on subsets of the domain. This type of incomplete data, also called functional fragments, is increasingly common in applications, e.g., Bugni (2012), Delaigle and Hall (2013), Liebl (2013), Goldberg et al. (2014), Kraus (2015), Delaigle and Hall (2016), Gromenko et al. (2017). The principal difficulty for inference with fragments is that temporal averaging is precluded by the incompleteness of the observed functions. Our formulation as an optimization problem enables to overcome this issue under certain assumptions because only averaging across individuals in the training data is needed, and not individual curves.

Since the observation domains may vary in the training sample and the new curve to classify may be observed on a different subset, it is natural to ask which domain should be used. We propose a domain selection strategy that looks for the best classifier with domain ranging from a minimum common domain to the entire domain of the function to classify. For various methods of selecting the best observation points we refer to, e.g., Ferraty et al. (2010), Delaigle et al. (2012),

Pini and Vantini (2016), Pini et al. (2017), Berrendero et al. (2017) and Stefanucci et al. (2018).

Our simulation study confirms that domain selection can result in a considerable reduction of the misclassification rate. Further simulations compare the performance of the three types of regularization. Among other findings, this study shows that the principal component and conjugate gradient classifiers often achieve comparable error rates but the latter usually needs a lower dimension of the regularization subspace, in agreement with a theoretical result we provide.

Application to a data set on the geometric features of the internal carotid artery in patients with and without aneurysm demonstrates the utility of the proposed methodology. These data consist of trajectories observed on intervals of different lengths. Previous analyses of these data used the common domain of all curves in classification. With our results we are able to include information beyond this minimum domain which leads to a substantial drop in the error rate of discrimination between risk groups.

General references on functional data analysis include Ramsay and Silverman (2005) and Horváth and Kokoszka (2012). Further relevant references are Cuesta-Albertos et al. (2007) for other methods based on one-dimensional projections, Berrendero et al. (2016) for variable selection in classification, Bongiorno and Goia (2016) and Dai et al. (2017b) for classification beyond the Gaussian setting, and Cuevas (2014) for an overview.

3.2 Regularized linear classification

3.2.1 Projection classifiers

We regard functional observations as random elements of the separable Hilbert space $L^2(\mathcal{I})$ of square integrable functions on a compact domain $\mathcal{I}$ equipped with inner product $\langle f, g\rangle = \int_{\mathcal{I}} f(t)g(t)dt$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. In most applications $\mathcal{I}$ is an interval and observations are curves but our results can be extended to other objects, such as surfaces or images. We consider classification of a Gaussian random function, X, into one of two groups of Gaussian random functions. Group 0 has mean μ_0, group 1 has mean μ_1. Both groups have covariance operator $\mathscr{R}$

defined as the integral operator

$$\mathscr{R}f(\cdot) = \int_{\mathcal{I}} \rho(\cdot, t) f(t) dt$$

with kernel $\rho(s,t) = \text{cov}\{X(s), X(t)\}$. In this section we assume that μ_0, μ_1 and $\mathscr{R}$ are known which corresponds to the asymptotic situation with an infinite training sample. To simplify the presentation we assume throughout the paper that the new observation to classify comes from both classes with equal prior probabilities. The general case is treated in the supplement.

Like Delaigle and Hall (2012a) we consider the class of centroid classifiers that are based on one-dimensional projections of the form $\langle X, \psi \rangle$, where ψ is a function in $L^2(\mathcal{I})$. If X belongs to group j, $j = 0, 1$, the distribution of $\langle X, \psi \rangle$ is normal with mean $\langle \mu_j, \psi \rangle$ and variance $\langle \psi, \mathscr{R}\psi \rangle$. Denote the corresponding Gaussian densities $f_{\psi,j}$. The optimal classifier based on $\langle X, \psi \rangle$ assigns X to the class $C_\psi(X)$ given by

$$C_\psi(X) = 1_{\{f_{\psi,1}(\langle X,\psi\rangle)/f_{\psi,0}(\langle X,\psi\rangle)>1\}} = 1_{\{\langle X-\mu_0,\psi\rangle^2 - \langle X-\mu_1,\psi\rangle^2>0\}} = 1_{\{T_\psi(X)>0\}},$$

where $T_\psi(X) = \langle X - \bar{\mu}, \psi \rangle \langle \mu, \psi \rangle$ with $\bar{\mu} = (\mu_0 + \mu_1)/2$ and $\mu = \mu_1 - \mu_0$. The misclassification probability of this classifier is

$$\begin{aligned} D(\psi) &= P_0\{C_\psi(X) = 1\}/2 + P_1\{C_\psi(X) = 0\}/2 \\ &= P_0(\langle X - \bar{\mu}, \psi \rangle \langle \mu, \psi \rangle > 0) \\ &= P_0(\langle X - \mu_0, \psi \rangle > |\langle \mu, \psi \rangle|/2) \\ &= 1 - \Phi\left(\frac{|\langle \mu, \psi \rangle|}{2\langle \psi, \mathscr{R}\psi \rangle^{1/2}} \right), \end{aligned} \tag{3.1}$$

where P_j, $j = 0, 1$ is the distribution of curves in group j and Φ is the standard normal cumulative distribution function.

To find the best function ψ, one would ideally like to maximize $|Q(\psi)|$, where

$$Q(\psi) = \frac{\langle \mu, \psi \rangle}{\langle \psi, \mathscr{R}\psi \rangle^{1/2}}.$$

Similarly to Delaigle and Hall (2012a) and Berrendero et al. (2017) we see that if $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then by the Cauchy–Schwarz inequality

$$\frac{|\langle \mu, \psi \rangle|}{\langle \psi, \mathscr{R}\psi \rangle^{1/2}} = \frac{|\langle \mathscr{R}^{-1/2}\mu, \mathscr{R}^{1/2}\psi \rangle|}{\langle \psi, \mathscr{R}\psi \rangle^{1/2}} \leq \frac{\|\mathscr{R}^{-1/2}\mu\| \|\mathscr{R}^{1/2}\psi\|}{\langle \psi, \mathscr{R}\psi \rangle^{1/2}} = \|\mathscr{R}^{-1/2}\mu\|. \tag{3.2}$$

If, moreover, $\|\mathscr{R}^{-1}\mu\| < \infty$, then the equality is achieved for $\psi = \mathscr{R}^{-1}\mu$. For this choice of ψ, or for any multiple of it, the probability of misclassification is $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$, which is positive due to the finiteness of the quantity $\|\mathscr{R}^{-1/2}\mu\|$ that can be seen as the signal-to-noise ratio. If $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then, regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$ or not, two Gaussian measures with mean difference μ and covariances $\mathscr{R}$ are mutually absolutely continuous and the value $1-\Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ is the Bayes error for classifying between them, i.e., the lowest possible misclassification probability for this problem not only among classifiers based on one-dimensional projections but among all possible classifiers, as explained by Berrendero et al. (2017). If $\|\mathscr{R}^{-1/2}\mu\| < \infty$ but $\|\mathscr{R}^{-1}\mu\| = \infty$, then the Bayes risk cannot be achieved by a projection classifier based on a bounded linear functional of the form $\langle X, \psi\rangle$ for some $\psi \in L^2(\mathcal{I})$. One can, however, use the theory of reproducing kernel Hilbert spaces to define a linear classifier that achieves the Bayes risk. We, however, do not pursue this idea because, as we see in the next subsection, approximations in the form of projections can asymptotically achieve the Bayes risk.

The maximization of $|Q(\psi)|$ can be solved as the task to

$$\text{maximize } \langle \mu, \psi\rangle \quad \text{subject to } \langle \psi, \mathscr{R}\psi\rangle = 1.$$

Using Lagrange multipliers $\langle \mu, \psi\rangle + \lambda(1 - \langle \psi, \mathscr{R}\psi\rangle)$ and taking Fréchet derivative with respect to ψ one obtains the equation $2\lambda\mathscr{R}\psi = \mu$. Solutions for all $\lambda > 0$, if they exist, i.e., if $\|\mathscr{R}^{-1}\mu\| < \infty$, yield the same optimal misclassification probability. Without loss of generality we take $\lambda = 1/2$. Thus the aim to minimize the error rate translates into the unconstrained quadratic optimization problem to maximize $\langle \mu, \psi\rangle - \frac{1}{2}\langle \psi, \mathscr{R}\psi\rangle$, or

$$\text{minimize } \tfrac{1}{2}\langle \psi, \mathscr{R}\psi\rangle - \langle \mu, \psi\rangle, \tag{3.3}$$

i.e., to the linear problem $\mathscr{R}\psi = \mu$.

3.2.2 Regularization

If $\psi = \mathscr{R}^{-1}\mu$ does not exist in $L^2(\mathcal{I})$, i.e., $\|\mathscr{R}^{-1}\mu\| = \infty$, there is no maximizer of $|Q(\psi)|$. One can instead consider an approximating, regularized problem that

can be solved. Regularization is typically used to solve ill-posed inverse problems, whose solution exists, in a stable way. There, the path of regularized solutions converges to the solution to the problem of interest. Here we are in a different situation in that no solution may exist. However, as we will see soon, paths of regularized solutions towards the possibly non-existent solution still turn out to be useful since the misclassification probability converges to the optimal value along these paths.

If a solution exists, one can approximate it by an iterative numerical method. This strategy can be applied also in situations where no solution exists. The idea is to construct a sequence of iterations of an appropriate numerical optimization method. The number of steps taken along this divergent sequence towards the non-existent solution can be seen as a regularization parameter. The conjugate gradient method is particularly suited for this situation.

The first K steps of the conjugate gradient method applied to the linear inverse problem $\mathscr{R}\psi = \mu$, or equivalently to the minimization of the quadratic functional $\frac{1}{2}\langle\psi, \mathscr{R}\psi\rangle - \langle\mu, \psi\rangle$, are described in Algorithm 1. This formulation of the algorithm is based on the multivariate version of (Phatak and de Hoog, 2002, Section 5) who give further references and also details on how applying the conjugate gradient method to the normal equations in linear regression leads to partial least squares regression. The functions ν_j are conjugate directions in the sense that $\langle\nu_j, \mathscr{R}\nu_k\rangle = 0$, $j \neq k$, and the functions ζ_j are called residuals in numerical analysis and are orthogonal, that is, $\langle\zeta_j, \zeta_k\rangle = 0$, $j \neq k$. In step j, the algorithm moves from the current approximate solution $\widehat{\psi}_j^{\mathtt{CG}}$ along the conjugate direction ν_j with step length h_j that minimizes the quadratic objective. The residual is then updated to ζ_{j+1}. The new conjugate direction ν_{j+1} is obtained by projecting the residual ζ_{j+1} onto the orthogonal complement of the span of the previous conjugate directions, where orthogonality is in the sense of the inner product $\langle\cdot, \mathscr{R}(\cdot)\rangle$.

The conjugate gradient approach is an example of dimension reduction regularization techniques. The method solves the minimization problem (3.3) with ψ restricted to the Krylov subspace $\mathtt{Kr}_{\mathtt{K}}(\mathscr{R}, \mu)$ spanned by $\mu, \mathscr{R}\mu, \ldots, \mathscr{R}^{K-1}\mu$, and also by the first K conjugate directions ν_j or the first K residuals ζ_j, i.e., it seeks to

$$\text{minimize } \tfrac{1}{2}\langle\psi, \mathscr{R}\psi\rangle - \langle\mu, \psi\rangle \quad \text{subject to } \psi \in \mathtt{Kr}_{\mathtt{K}}(\mathscr{R}, \mu).$$

Initialize $\psi_0^{\mathtt{CG}} = 0$, $\nu_0 = \zeta_0 = \mu$
Repeat for $j = 0, \ldots, K-1$
$h_j = \langle \nu_j, \zeta_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\psi_{j+1}^{\mathtt{CG}} = \psi_j^{\mathtt{CG}} + h_j \nu_j$
$\zeta_{j+1} = \mu - \mathscr{R}\psi_{j+1}^{\mathtt{CG}}$ $(= \zeta_j - h_j \mathscr{R}\nu_j)$
$g_j = -\langle \zeta_{j+1}, \mathscr{R}\nu_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\nu_{j+1} = \zeta_{j+1} + g_j \nu_j$
Output $\psi_K^{\mathtt{CG}}$

Algorithm 1: Conjugate gradient regularized classification direction

The projection direction that solves this minimization is $\psi_K^{\mathtt{CG}}$.

Another popular choice is to

$$\text{minimize } \tfrac{1}{2}\langle \psi, \mathscr{R}\psi \rangle - \langle \mu, \psi \rangle \quad \text{subject to } \psi \in \mathtt{E_K}(\mathscr{R}),$$

where $\mathtt{E_K}(\mathscr{R})$ is the subspace spanned by the first K eigenfunctions, $\varphi_1, \ldots, \varphi_K$, of $\mathscr{R}$ in the spectral decomposition

$$\mathscr{R} = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ are the eigenvalues. The solution $\psi_K^{\mathtt{PC}} = \sum_{j=1}^{m} \lambda_j^{-1} \langle \mu, \varphi_j \rangle \varphi_j$ gives the principal component classifier of Delaigle and Hall (2012a).

In general one can

$$\text{minimize } \tfrac{1}{2}\langle \psi, \mathscr{R}\psi \rangle - \langle \mu, \psi \rangle \quad \text{subject to } \psi \in \mathtt{S_K},$$

where $\mathtt{S_K}$ is the K-dimensional subspace generated by some functions $s_1, \ldots, s_K$ such that s_j, $j = 1, 2, \ldots$ generate the range of $\mathscr{R}$. Let $\mathscr{P}_K$ be the projection operator that projects on $\mathtt{S_K}$, $\mathscr{R}_K = \mathscr{P}_K \mathscr{R} \mathscr{P}_K$ and $\mathscr{R}_K^- = \mathscr{P}_K \mathscr{R}^{-1} \mathscr{P}_K$. Then the solution of the regularized minimization problem is $\psi_K = \mathscr{R}_K^- \mu$. More explicitly, considering solutions of the form $\psi_K = \sum_{j=1}^{K} c_j s_j$ leads to the K-variate minimization of $\frac{1}{2} c^{\mathsf{T}} Q c - u^{\mathsf{T}} c$ with the matrix Q with $Q_{jk} = \langle s_j, \mathscr{R} s_k \rangle$ and vector u with $u_j = \langle \mu, s_j \rangle$, i.e., to the solution with coefficients $c = Q^{-1} u$. In the case of the Krylov subspace, the iterative conjugate gradient method given in Algorithm 1 is, however, preferred because the matrix Q is ill-conditioned.

We can also use another approach to regularization based on ridge regression. Optimizing the misclassification probability in a ball with radius $\theta^{1/2}$ leads to the task to

$$\text{minimize } \tfrac{1}{2}\langle\psi, \mathscr{R}\psi\rangle - \langle\mu, \psi\rangle \quad \text{subject to } \|\psi\|^2 \leq \theta,$$

or, equivalently,

$$\text{minimize } \tfrac{1}{2}\langle\psi, \mathscr{R}\psi\rangle - \langle\mu, \psi\rangle + \tfrac{1}{2}\alpha\|\psi\|^2,$$

where $\alpha \geq 0$ is a regularization parameter. The solution is $\psi_\alpha^{\mathtt{R}} = \mathscr{R}_\alpha^{-1}\mu$, where $\mathscr{R}_\alpha = \mathscr{R} + \alpha\mathscr{I}$ and $\mathscr{I}$ is the identity operator. Despite its practical performance and simplicity of theoretical analysis the functional ridge classifier does not seem to have been considered before.

There is an important difference between the conjugate gradient method and the other approaches. While principal components and the ridge method regularize the problem without the main goal in mind, the conjugate gradient approach greedily follows the goal of optimal classification. Indeed, the conjugate gradient method as an iterative optimization procedure constructs the regularization path focusing on the minimization of the misclassification probability whereas the other approaches regularize by modifying the operator to be inverted regardless of the goal.

From the computational point of view the conjugate gradient method is the simplest one because it does not require inversion or eigendecomposition.

3.2.3 Properties of regularization paths

While ψ_K, the solution regularized by a subspace constraint, in general need not converge as $m \to \infty$ since no solution to the unconstrained minimization problem may exist, the misclassification probability associated with the linear classifier given by ψ_K converges along the regularization path. The following and all other results are proved in the appendix.

Proposition 1. *The misclassification probability of the regularized linear classifier based on $\psi_K = \mathscr{R}_K^- \mu$ converges to $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ as $m \to \infty$.*

The above result holds regardless of whether the unconstrained minimization problem (3.3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is either positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$, or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$. As discussed earlier, the optimal error is achieved exactly by the one-dimensional projection on $\psi = \mathscr{R}^{-1}\mu$, when $\|\mathscr{R}^{-1}\mu\| < \infty$. Even when $\|\mathscr{R}^{-1}\mu\| = \infty$, both dimension reduction techniques, conjugate gradients and principal components, and also ridge regularization as we will see soon, achieve the optimal limiting error rate along a possibly non-convergent path of one-dimensional projection directions.

It is natural to investigate and compare how quickly the misclassification rate approaches the limit for both main types of subspace regularization. It turns out that the conjugate gradient classifier, being a greedy, goal-oriented procedure, performs better than or at least equally well as the principal component classifier with the same dimension.

Proposition 2. *Regardless of whether the optimal misclassification probability can be achieved exactly or along a regularization path, i.e., $\|\mathscr{R}^{-1}\mu\| < \infty$ or $\|\mathscr{R}^{-1}\mu\| = \infty$, and regardless of whether the optimal misclassification probability is zero or positive, i.e., $\|\mathscr{R}^{-1/2}\mu\| = \infty$ or $\|\mathscr{R}^{-1/2}\mu\| < \infty$, the misclassification probability of the principal component classifier using K components is higher than or equal to the misclassification probability of the K-step conjugate gradient classifier.*

(Phatak and de Hoog, 2002, Subsection 6.2) showed in the multivariate setting that "PLS fits closer than PCR." In infinite dimension in the context of kernel partial least squares (Blanchard and Krämer, 2010, Theorem 1) showed that the partial least squares solution is closer to the true solution of the inverse problem than the principal component solution with the same number of components. Unlike these results, our result in Proposition 2 does not assume the existence of a solution and instead focuses on the values of the misclassification probability.

Although Proposition 2 suggests that the conjugate gradient method will typically use less components than the principal component method to achieve the best result, it does not necessarily mean that the resulting misclassification probability with the best number of components will be better for conjugate gradients. We

address this question in the simulation study. A similar phenomenon in the context of regression was previously studied in the literature on partial least squares in finite dimension and recently in the functional setting by Febrero-Bande et al. (2017).

Similarly to the case of subspace regularization we obtain below the convergence of the error probability of the ridge classifier. The result holds regardless of whether the unconstrained minimization problem (3.3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is either positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$, or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$.

Proposition 3. *The misclassification probability of the regularized linear classifier based on* $\psi_\alpha^{\mathtt{R}} = \mathscr{R}_\alpha^{-1}\mu$ *converges to* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ *as* $\alpha \to 0+$.

3.3 Empirical classifiers for fragmentary functions

3.3.1 Construction of classifiers with incomplete training samples

So far we discussed classification assuming that the parameters of each group are known. We now present the empirical version with unknown distributional parameters but with a finite training data set available, and show that under some regularity conditions such classifiers can achieve asymptotically the same optimal error rate as if there were infinite training samples. We aim to do this not only in the case of fully observed functions but also in the case of incomplete curves. Incompleteness can occur in the training data, with each curve possibly observed on a different domain, and in the new curve we wish to classify. One strategy would be to consider all curves on the intersection of their observation domains, if it is non-empty. However, such a restriction can be too severe and is not necessary. We will construct classifiers that use the observed new curve on a set $\mathcal{I}$ which may be its entire observation set or a subset of it without requiring that all training curves be completely observed on $\mathcal{I}$.

For group j let there be a training sample consisting of N_j curves $X_{j1}, \ldots, X_{jN_j}$. The training data are assumed to be mutually independent. Curves may be observed incompletely with values known only on a subset O_{ji} of the domain and no

information about the values on the complement. The observation domains are assumed to be independent of the curves and consist of a finite union of intervals. By $O_{ji}(t)$ we denote the indicator that the curve X_{ji} is observed at time t. Similarly, let $U_{ji}(s,t)$ indicate observation at times s and t, i.e., $U_{ji}(s,t) = O_{ji}(s)O_{ji}(t)$.

The mean μ_j in group $j = 0, 1$ can be estimated by the cross-sectional average

$$\widehat{\mu}_j(t) = \frac{1_{[\tilde{N}_j(t)>0]}}{\tilde{N}_j(t)} \sum_{i=1}^{n_j} O_{ji}(t) X_{ji}(t),$$

where $\tilde{N}_j(t) = \sum_{i=1}^{n_j} O_{ji}(t)$ is the total number of observed curves in group j at time t. The covariance kernel $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations of groupwise centred curves. Formally, the estimator is

$$\widehat{\rho}(s,t) = \frac{M_1(s,t)\widehat{\rho}_1(s,t) + M_2(s,t)\widehat{\rho}_2(s,t)}{M_1(s,t) + M_2(s,t)},$$

where

$$\widehat{\rho}_j(s,t) = \frac{1_{[M_j(s,t)>0]}}{M_j(s,t)} \sum_{i=1}^{n_j} U_{ji}(s,t)\{X_{ji}(s) - \widehat{\mu}_{jst}(s)\}\{X_{ji}(t) - \widehat{\mu}_{jst}(t)\}$$

and $M_j(s,t) = \sum_{i=1}^{n_j} U_{ji}(s,t)$ and $\widehat{\mu}_{jst}(s) = \frac{1_{[M_j(s,t)>0]}}{M_j(s,t)} \sum_{i=1}^{n_j} U_{ji}(s,t) X_{ji}(s)$. If $\tilde{N}_j(t) = 0$ or $M_j(s,t) = 0$ in the above definitions, the estimators are defined as $\widehat{\mu}_j(t) = 0$ or $\widehat{\rho}_j(s,t) = 0$, respectively. This happens with asymptotically vanishing probability under Assumption 1 below.

Let the new, independent curve to classify, $X_{\mathtt{new}}$, be observed on the domain $O_{\mathtt{new}}$. Let us fix the target domain $\mathcal{I} \subseteq O_{\mathtt{new}}$ on which we aim to apply the classifier to $X_{\mathtt{new}}$. The empirical classifier $\widehat{C}_{\widehat{\psi}}$ trained on partially observed curves is defined like the theoretical one with unknown quantities replaced by their estimators. It assigns $X_{\mathtt{new}}$ restricted to $\mathcal{I}$ to the class $\widehat{C}_{\widehat{\psi}}(X_{\mathtt{new}}) = 1_{[\widehat{T}_{\widehat{\psi}}(X_{\mathtt{new}})>0]}$, where $\widehat{T}_{\widehat{\psi}}(X_{\mathtt{new}}) = \langle X_{\mathtt{new}} - \tilde{\mu}, \widehat{\psi}\rangle\langle\widehat{\mu}, \widehat{\psi}\rangle$. Here $\tilde{\mu} = (\widehat{\mu}_0 + \widehat{\mu}_1)/2$ and $\widehat{\mu} = \widehat{\mu}_1 - \widehat{\mu}_0$ with $\widehat{\mu}_j$ being the estimators defined above, restricted to $\mathcal{I}$. The projection direction $\widehat{\psi}$ is one of $\widehat{\psi}_K^{\mathtt{CG}}$, $\widehat{\psi}_K^{\mathtt{PC}}$ or $\widehat{\psi}_\alpha^{\mathtt{R}}$ constructed by conjugate gradient, principal component or ridge regularization applied to $\widehat{\mu}$ and $\widehat{\mathscr{R}}$, with $\widehat{\mathscr{R}}$ being the integral operator with kernel $\widehat{\rho}(s,t)$ introduced above, restricted to $\mathcal{I} \times \mathcal{I}$.

It is an important feature of all methods discussed in the previous section that they are formulated in terms of the population parameters, i.e., mean difference and covariance operator, and not in terms of individual observations in the training set. The population parameters can be consistently estimated by averaging individual observations whereas temporal averaging of individual curves, e.g., in inner products, is impossible due the incompleteness of the observed functions. In particular, the conjugate gradient method can be applied to fragmentary training data whereas usual algorithms for multivariate or functional partial least squares, e.g., De Jong (1993), (Hastie et al., 2009, Algorithm 3.3) and (Delaigle and Hall, 2012b, Subsection 4.2, Appendix A.2), involve the computation of certain scores, i.e., inner products, for individual curves.

3.3.2 Asymptotic behaviour along the empirical regularization path

We aim to study the behaviour of classifiers for incomplete training samples of increasing size with decreasing amount of regularization. Previous asymptotic results in related settings include those of Delaigle and Hall (2013) who establish the consistency of empirical principal component classifiers based on partially observed training data. In the setting of complete curves Berrendero et al. (2017) use dimension reduction regularization by evaluation of curves at a finite set of arguments. They show the consistency of the empirical version but do not study asymptotics for decreasing amount of regularization, i.e., do not let the dimension grow. Baíllo et al. (2011a) study optimal classifiers for Gaussian measures based on Radon–Nikodym derivatives and investigate the performance of their empirical version in the special class of processes with triangular covariance functions. In contrast, all our methods, including the ridge approach not considered previously, are developed for fragmentary training samples and shown to achieve the Bayes error rate for general Gaussian processes along the empirical regularization path, as we now explain.

The following assumptions will be needed for the derivation of asymptotic properties of empirically trained regularized linear classifiers.

Assumption 1.

(a) Let the distributions in groups $j = 0, 1$ satisfy $\mathbb{E}_{P_j}(\|X\|^4) < \infty$.

(b) For a domain $\mathcal{I}$ let there be $\delta > 0$ such that the observation patterns in training samples $j = 0, 1$ satisfy, as $N_j \to \infty$,

$$\sup_{(s,t)\in\mathcal{I}\times\mathcal{I}} P\{N_j^{-1} M_j(s,t) > \delta\} = O(N_j^{-2}).$$

Assumption (a) is the standard assumption that guarantees the consistency of the empirical mean and covariance operator for samples of completely observed curves; see, e.g., Bosq (2000) or Horváth and Kokoszka (2012). It was shown in (Kraus, 2015, Proposition 1) under the additional assumption (b) with $\mathcal{I}$ equal to the entire domain of the curves that the root-n consistency of the sample mean and covariance restricted to $\mathcal{I}$ continues to hold in the fragmentary setting. In particular, it follows that $\|\widehat{\mu}_j - \mu_j\| = O_P(N_j^{-1/2})$, and hence $\|\widehat{\mu} - \mu\| = O_P(N^{-1/2})$ for $N = \min(N_0, N_1) \to \infty$, and also $\|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty = O_P((N_0 + N_1)^{-1/2})$, where $\|\cdot\|_\infty$ is the operator norm. When $\mathcal{I}$ is a subset of the domain, analogous results hold for the obvious restrictions of the functions and integral kernels to $\mathcal{I}$. Assumption (b) means that at all pairs of time points there is an asymptotically non-negligible fraction of observed values. Assumption (b) is less restrictive than the requirement that there be complete curves in the sample. It may be satisfied, for example, in situations where observed curves consist of several shorter fragments. If the assumption is not satisfied because the data contain only one short fragment per curve, other estimation methods can be used, see, e.g., Delaigle and Hall (2016) and Descary and Panaretos (`https://arxiv.org/abs/1708.02491`)

We now study the asymptotic behaviour of the empirical classifier when the number m_n of steps of the conjugate gradient algorithm grows as the training sample size grows. We establish under certain conditions on the regularization path the convergence of the misclassification probability of the conjugate gradient classifier trained on collections of functional fragments to the same optimal limit as for the theoretical conjugate gradient classifier with infinite training sample, regardless of whether the limiting error rate is zero or positive and regardless of whether the limit can be theoretically achieved exactly or along the path.

Theorem 1. *Let Assumption 1 hold. Assume that* $N = \min(N_0, N_1) \to \infty$ *and* $K_N \to \infty$ *in such a way that* $K_N \leq CN^{1/2}$ *for some* $C > 0$ *and*

$$N^{-1/2}\omega_{K_N}^{-1}\|\gamma^{(K_N)}\| + N^{-1}\omega_{K_N}^{-3} \to 0, \tag{3.4}$$

where ω_{K_N} *is the smallest eigenvalue of the* $(K_N \times K_N)$*-matrix* H *with entries* $h_{jk} = \langle \kappa_j, \mathscr{R}\kappa_k \rangle$ *for* $\kappa_j = \mathscr{R}^{j-1}\mu$ *and the* K_N*-vector* $\gamma^{(K_N)}$ *is defined as* $\gamma^{(K_N)} = H^{-1}d$ *with* d *being the* m_n*-vector with components* $d_j = \langle \mu, \kappa_j \rangle$*. Then the misclassification probability of the empirical regularized linear classifier based on* $\widehat{\psi}_{K_N}^{\mathtt{CG}}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$*.*

Condition (3.4) guarantees that the number of components does not grow too fast in relation to the growing number of training observations and to the increasingly ill conditioning of the theoretical problem. Condition (3.4) is analogous to (5.10) in Delaigle and Hall (2012b) for partial least squares. The vector $\gamma^{(K_N)}$ contains the coefficients of the theoretical regularized solution $\psi_{K_N}^{\mathtt{CG}}$ with respect to the non-orthogonal basis $\kappa_1, \ldots, \kappa_{K_N}$ of the Krylov subspace $\mathtt{Kr}_{\mathtt{K_N}}(\mathscr{R}, \mu)$, i.e., $\psi_{K_N} = \sum_{j=1}^{K_N} \gamma_j^{(K_N)} \kappa_j$. The eigenvalues of H are called the Ritz values in numerical analysis. For details on connections with partial least squares see Lingjærde and Christophersen (2000).

In the proof in Subsection B.1.4 we make use of the results of Delaigle and Hall (2012b) on the consistency of partial least squares regression for functional data. These results were obtained for situations that are different from our setting in several ways. In particular, we work with functional fragments instead of complete curves, the conjugate gradient path differs from partial least squares regression, e.g., in the group centring in the estimation of the covariance, and we do not require that the population inverse problem, $\mathscr{R}\psi = \mu$ in our context, have a solution. However, our inspection of the underlying technical arguments in Delaigle and Hall (2012b) showed that appropriate analogous results can be obtained and used in our setting, as we explain in the proof.

Next, we show that the empirically trained principal component classifier with increasing number of components asymptotically achieves the optimal misclassification probability.

Theorem 2. *Let Assumption 1 hold. Assume that* $N = \min(N_0, N_1) \to \infty$ *and* $K_N \to \infty$ *in such a way that* $\lambda_{K_N}^4 N \to \infty$ *and* $\lambda_{K_N}^2 N(\sum_{j=1}^{K_N} a_j)^{-2} \to \infty$*, where* $a_1 = 2^{3/2}(\lambda_1 - \lambda_2)^{-1}$ *and* $a_j = 2^{3/2}\max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\}$*,* $j = 2, 3, \ldots$ *Then the misclassification probability of the empirical regularized linear classifier based on* $\widehat{\psi}_{K_N}^{\mathtt{PC}}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$*.*

The conditions on the principal component regularization path are the same as in the case of functional principal component regression (Cardot et al., 1999), (Cardot et al., 2007). Unlike in the functional linear model it is not assumed that the inverse problem has a solution since the goal is not to estimate the possibly non-existent bounded linear regression functional.

Finally, the empirical ridge classifier with finite training data asymptotically attains the same optimal error rate as its theoretical counterpart. Unlike for conjugate gradients and principal components, the conditions on the ridge path do not involve parameters of the data distributions because no subspace is constructed.

Theorem 3. *Let Assumption 1 hold. Assume that* $N = \min(N_0, N_1) \to \infty$ *and* $\alpha_N \to 0+$ *in such a way that* $\alpha_N^4 N \to \infty$*. Then the misclassification probability of the empirical regularized linear classifier based on* $\widehat{\psi}_{\alpha_N}^{\mathtt{R}}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$*.*

3.3.3 Selection of the regularization parameter

The regularization parameter can be selected by minimizing an estimate of the misclassification probability. We use leave-one-out cross-validation. Algorithm S1 in the supplementary document provides details of cross-validation in the presence of incomplete curves. The best value of the regularization parameter is searched for over a grid of values, e.g., the values corresponding to integer degrees of freedom up to some maximum value. The degrees of freedom for the subspace methods are the dimension of the subspace and for the ridge method they are defined as the trace of $(\widehat{\mathscr{R}}+\alpha\mathscr{I})^{-1}\widehat{\mathscr{R}}$, that is, $\sum_{j=1}^{N_0+N_1} \widehat{\lambda}_j/(\widehat{\lambda}_j+\alpha)$, where $\widehat{\lambda}_j$ are the eigenvalues of $\widehat{\mathscr{R}}$. The maximum number of degrees of freedom we use is one fifth of the number of curves.

3.4 Domain selection

To classify the new curve $X_{\mathtt{new}}$ observed on $O_{\mathtt{new}}$, we apply the classifier on the target domain $\mathcal{I} \subseteq O_{\mathtt{new}}$. We now consider the choice of the target domain. One extreme possibility would be to restrict attention to the intersection of the observation domains of all curves, say $\mathcal{I}_{\mathtt{min}}$, if it is non-empty. An obvious drawback of this approach is that one can lose discrimination power because the difference between the classes may be more pronounced outside $\mathcal{I}_{\mathtt{min}}$. An advantage of our methodology is its capability to work with incomplete curves since the empirical construction of the projection direction only requires the estimation of μ and $\mathscr{R}$ on the target domain. Hence one may look at a larger domain than $\mathcal{I}_{\mathtt{min}}$. A natural choice is the largest subset of $O_{\mathtt{new}}$ containing enough data for the estimation of the classifier, i.e., satisfying Assumption 1(b), and enough functions for validation in the cross-validation procedure, i.e., with sufficiently large set V in Algorithm S1 in the supplement. This way one hopes to capture the widest range of shapes of the group difference. On the other hand, not even this maximal domain, say $\mathcal{I}_{\mathtt{max}}$, may lead to the best classification accuracy because one includes more uncertainty in the estimation due to missing values and the mean difference may not be important in the added part of the domain. Therefore, it seems reasonable to look also at intermediate choices between $\mathcal{I}_{\mathtt{min}}$ and $\mathcal{I}_{\mathtt{max}}$.

Here we present a domain selection strategy for the most common case of interval observation sets. The idea, worked out in detail in Stefanucci et al. (2018), is to construct the classifier on a series of intervals, ranging from the common domain $\mathcal{I}_{\mathtt{min}}$ to the maximal domain $\mathcal{I}_{\mathtt{max}}$, extending step by step the working interval by a fixed percentage. More formally, we consider a sequence of nested intervals $\mathcal{I}_{\mathtt{min}} = \mathcal{I}_0 \subset \mathcal{I}_1 \subset \cdots \subset \mathcal{I}_\ell \subset \cdots \subset \mathcal{I}_L = \mathcal{I}_{\mathtt{max}}$ starting from $\mathcal{I}_{\mathtt{min}}$ and ending in $\mathcal{I}_L = \mathcal{I}_{\mathtt{max}}$ and build the classifier on each of them. The regularization parameter for the $\ell^{\mathtt{th}}$ domain is selected by cross-validation as described in the supplement. Among these $L + 1$ candidates we select the one that minimizes the cross-validation estimate of error.

The search strategy can be extended by considering larger systems of candidate domains, e.g., one can vary both endpoints independently. The idea can be generalized to other situations, e.g., non-interval observation sets, multivariate functional

data or functions indexed by multivariate arguments. In each situation one needs to define a meaningful system of domains and optimize the cross-validation score over it.

3.5 Simulations

3.5.1 Behaviour of regularized classifiers on complete data

In this section we illustrate the behaviour of the three estimators of ψ under different settings. We consider Gaussian processes on $[0, 1]$ with covariance kernel $\rho(s, t) = \exp(-|s - t|^2/0.01)$ and mean function depending on the group label. Group 0 has mean $\mu_0(t) = 0$ in each setting. Group 1 has mean $\mu_1(t) = \mu(t)$ for which we consider eight different forms, (i) ct, (ii) $c(t - 0.5)^2$, (iii) $c(t - 0.5)^3$, (iv) $c\sin(20t)$, (v) $c\varphi_1(t)$, (vi) $c\varphi_{10}(t)$, (vii) $cb(t; 5, 5)$, (viii) $cb(t; 2, 6)$, where φ_j is the j^{th} eigenfunction of the kernel ρ and $b(t; \alpha, \beta) = t^{\alpha-1}(1 - t)^{\beta-1}$. The parameter c is selected in each case to yield a reasonable misclassification rate.

In each of 5000 repetitions we generated 50 curves from each group and evaluated them on a grid of 100 equispaced points in $[0, 1]$. We also generated a new observation that could arise from group 0 or group 1 with equal probability. Then we constructed the regularized classification direction by the principal component, conjugate gradient and ridge method with K degrees of freedom and predicted the label of the new observation. We considered $K = 1, \ldots, 20$, corresponding to a reasonable minimum of five observations per degree of freedom.

The results are plotted in Fig. 3.1. It shows the misclassification proportion over the 5000 repetitions as a function of K for different choices of $\mu(t)$. As expected, the conjugate gradient method performs well in all settings and is not much affected by the particular shape of $\mu(t)$. By contrast, the performance of the principal component classifier strongly depends on $\mu(t)$. To see this, consider two extreme situations in settings (v) and (vi). The classification error of principal components is close to the one of conjugate gradients in case (v), where $\mu(t)$ is the first eigenfunction, but it is much higher at lower dimensions in case (vi), where $\mu(t)$ is the tenth eigenfunction. In the latter case, the principal component method reaches the same level of error as the conjugate gradient method only when $m = 10$

or more. These findings agree with the theoretical result of Proposition 2 and also with conclusions of Delaigle and Hall (2012a) and Febrero-Bande et al. (2017) who point out that principal components need more degrees of freedom than partial least squares to reach good performance. In this regard ridge regularization seems to be between the two subspace methodologies. It is more similar to conjugate gradients in most cases. In particular in case (vi) it does not completely fail at low degrees of freedom because it does not construct a subspace that can possibly miss the important information. On the other hand it also suffers in this situation, where $\mu(t)$ is on the tail of the spectrum, because ridge penalization shrinks higher index spectral components more than lower index components. However, with sufficiently many degrees of freedom differences fade away and the three methods behave similarly.

Additional simulation results are reported in the supplementary document. They show that similar conclusion can be drawn when functions have non-smooth trajectories and that the capability to discriminate between two groups with different means is robust with respect to the assumption of equal covariances. Results for increased training sample size are also provided.

3.5.2 Performance of cross-validation for selection of degrees of freedom

We employed simulations to investigate the performance of leave-one-out cross validation in choosing the right level of regularization. The settings were the same as before but classification was done using the number of degrees of freedom selected by leave-one-out cross-validation. We summarize the classification error in Table 3.1. A general observation is that cross-validation performs well as a selector of the best amount of regularization since the value of misclassification rate in Table 3.1 is in each case close to the corresponding minimum error in Fig. 3.1. Principal components appear to perform worst while the conjugate gradient and ridge methods have comparable performance. The latter two methods nearly achieve the respective minimum error. Table 3.2 reports the mean and median selected degrees of freedom. We see that the principal component method often uses considerably more degrees of freedom than the other methods. This is particularly

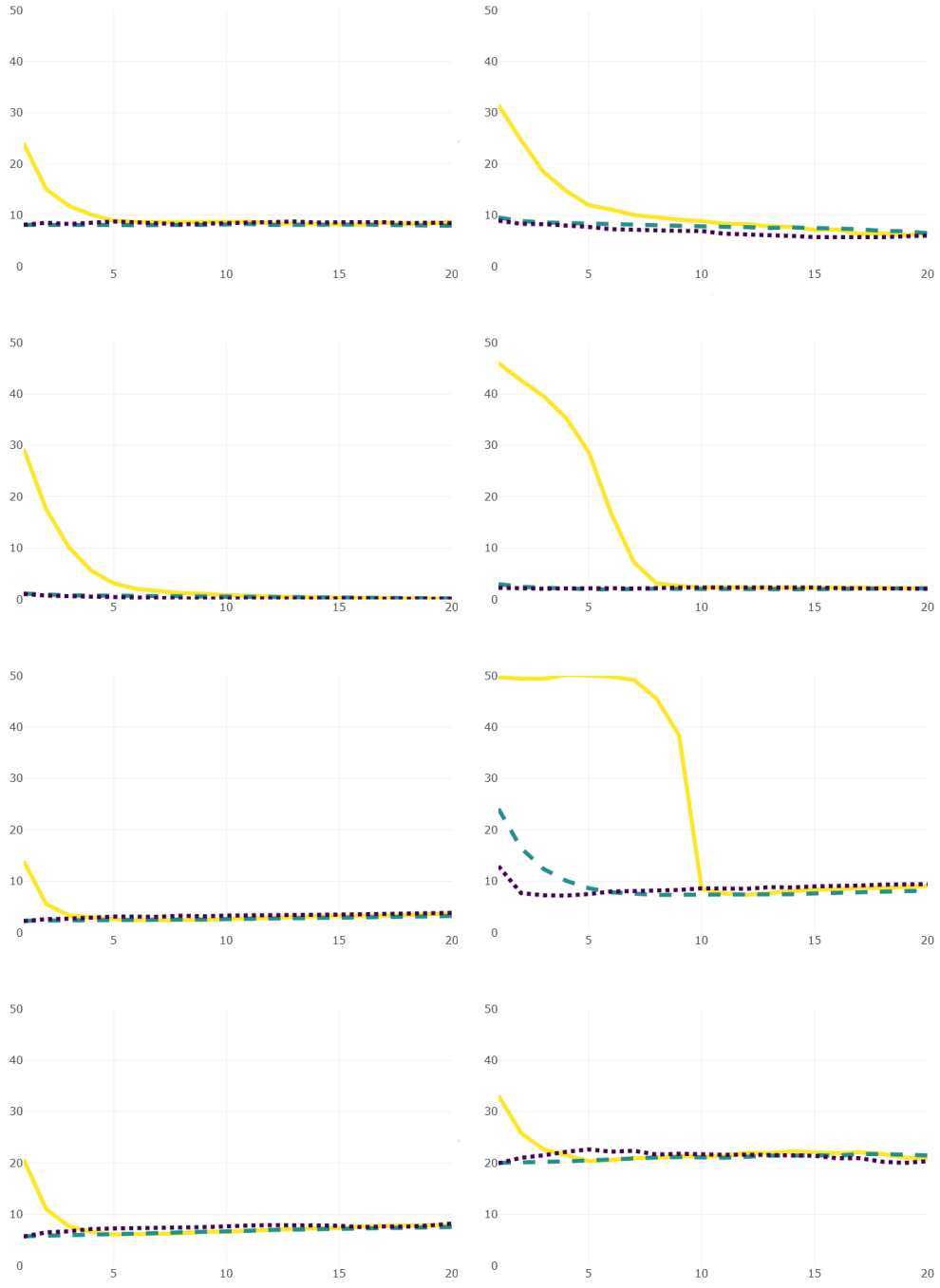

Figure 3.1: *Misclassification rate (%) versus degrees of freedom for different forms of $\mu(t)$, (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, (viii) asymmetric beta, for principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.*

Table 3.1: *Misclassification rate (%) and its standard error achieved by classifiers with degrees of freedom selected by cross-validation for different settings, with minimum missclassification rate on the second row for each classifier*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC	13.0 (0.34)	8.3 (0.28)	1.3 (0.11)	2.5 (0.16)	7.2 (0.26)	7.6 (0.27)	10.7 (0.31)	26.2 (0.44)
	8.1	6.1	0.1	2.2	2.4	7.4	6.1	20.4
CG	8.6 (0.28)	6.5 (0.25)	0.7 (0.09)	2.1 (0.14)	2.6 (0.16)	7.8 (0.27)	6.1 (0.24)	20.9 (0.41)
	8.1	5.7	0.1	2.1	2.2	7.2	5.7	19.9
R	8.4 (0.28)	7.7 (0.27)	0.7 (0.09)	2.2 (0.15)	2.4 (0.15)	7.9 (0.27)	6.1 (0.24)	20.8 (0.41)
	7.9	6.5	0.2	2.0	2.3	7.3	5.7	20.0

PC, principal components; CG, conjugate gradients; R, ridge.

Table 3.2: *Mean (median) degrees of freedom selected by cross-validation*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
Principal components	8.2 (7)	14.3 (15)	9.9 (9)	10.9 (10)	4.6 (4)	11.9 (11)	5.3 (4)	8.6 (6)
Conjugate gradients	5.4 (3)	10.7 (11)	3.4 (2)	4.5 (2)	2.4 (1)	4.9 (3)	2.7 (1)	8.6 (7)
Ridge	6.4 (3)	11.6 (13)	6.0 (3)	6.1 (4)	2.7 (1)	9.3 (8)	3.4 (1)	6.7 (3)

interesting in case (v), where the mean difference equals the first eigenfunction and thus one component should be the best choice in theory. These results once again document the general phenomenon that principal components are not appropriate for inference about means due to the possible lack of informativeness of the principal components about the mean and the extra uncertainty associated with the estimation of these components.

3.5.3 Missing data and domain extension

We now show the usefulness of domain extension presented in Section 3.4. We considered Gaussian processes on $[0, 1]$ with the same covariance as before and with three scenarios for the mean difference of the form of a multiple of a beta density. These were (a) $b(t; 2, 6)$, (b) $b(t; 5, 5)$ and (c) $b(t; 6, 2)$ which reflect situations in which discrimination due to a peak is in the left, central or right part of the domain, respectively. We sampled 50 curves from each group 0 on a sequence of 100 equispaced points in $[0, 1]$. Then we generated endpoints of the observation interval for each curve from the uniform distribution on $(0.5, 1)$, that is, each curve was observed between 0 and the endpoint and missing beyond the endpoint. Also the new observation had an endpoint sampled between 0.5 and 1. So the first

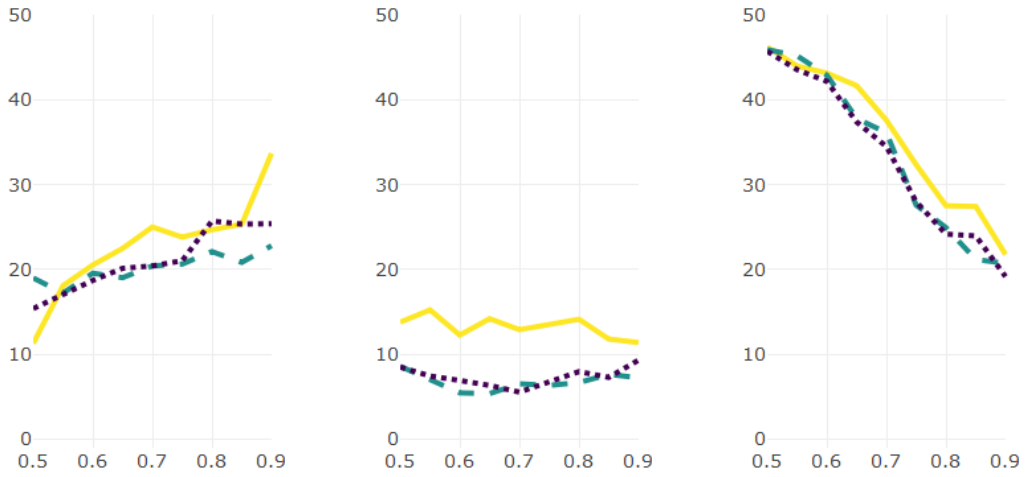

Figure 3.2: *Misclassification rate (%) as a function of the domain extension for $\mu(t)$ being the Beta(2,6) (left), Beta(5,5) (middle), Beta(6,2) (right) density for principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers with selected degrees of freedom. Classification is performed on the domains $[0, u]$, $u \in [0.5, 0.9]$, error values are plotted against u.*

half of $[0, 1]$, $\mathcal{I}_{\texttt{min}} = [0, 0.5]$, was the common observation domain of all curves. We considered extensions of $\mathcal{I}_{\texttt{min}}$ to $\mathcal{I}_\ell = [0, 0.5 + 0.05\ell]$, $\ell = 0, \ldots, 8$. For each interval of this form that was contained in the observation domain of the curve to classify we estimated the classifiers choosing the best degrees of freedom via cross-validation and classified the new curve. This was repeated 1000 times. We show the behaviour of the resulting classification error as a function of the endpoint of the extended domain in Fig. 3.2.

In the case, where the peak of the mean difference is in the left part of $[0, 1]$, extending the domain does not lead to better classification. In this case the interval, where discrimination is, corresponds to the part of the domain where all the data are available and inflating the domain only incurs uncertainty due to the presence of missing data. In the second case the peak of the mean difference is exactly at 0.5 and extending the domain leads to little improvement. The last case is the converse of the first one, the discrimination is mainly in the right part of $[0, 1]$. Here extending the domain considerably reduces the error because good

Table 3.3: *Misclassification rate (%) and its standard error achieved by classifiers with domain and degrees of freedom selected by cross-validation for different settings, with minimum and maximum missclassification rate in square brackets*

	(a)	(b)	(c)
Principal components	18.1 (0.38) [11.3,33.7]	11.9 (0.32) [11.4,15.2]	31.1 (0.46) [21.8,46.0]
Conjugate gradients	19.6 (0.39) [15.4,25.7]	7.4 (0.26) [5.6,9.3]	30.4 (0.46) [19.2,45.7]
Ridge	22.4 (0.42) [17.2,22.8]	6.9 (0.25) [5.4,8.6]	28.4 (0.45) [20.7,45.9]

classification is only possible by employing the right part of the domain. The classification error is about 45 % using only $\mathcal{I}_{\texttt{min}}$ but drops to about 20 % using also part of the interval where the data are partially observed.

3.5.4 Performance with selected domain

We have seen that domain extension may or may not lead to an improvement of the performance of classifiers, depending on the interplay between the form of the mean difference, the covariance structure and the missingness pattern. In practice, one is not an oracle with access to misclassification errors for candidate subsets whose estimates are plotted in Fig. 3.2 and hence selects the best domain by cross-validation. In Table 3.3 we report simulation results for classifiers with selected domain and degrees of freedom for the same configurations as in Subsection 3.5.3. Selection of the domain leads to a considerable improvement of the error rate in comparison with the worst performing domain. On the other hand, this improvement has some limitations and there remains a gap between the achieved and the best value. This can be explained by the fact that cross-validation provides only an estimate of the error and not the true value.

3.6 AneuRisk data example

We apply the proposed methodology to the AneuRisk dataset from an interdisciplinary project that aimed at investigating the role of vessel morphology, blood fluid dynamics and biomechanical properties of the vascular wall on the pathogenesis of cerebral aneurysms. See Sangalli et al. (2014b) for an introduction to the data. This dataset has been previously analyzed in several works with different methodological focuses, such as function and derivative estimation (Sangalli et al.,

2009b), exploratory analysis and classification (Sangalli et al., 2009a), alignment and clustering (Sangalli et al., 2014a), (Sangalli et al., 2010) among others.

The data consist of measurements of the radius and curvature of the internal carotid artery in a sample of 65 patients of which 33 have an aneurysm at the bifurcation of the vessel or after it, while the other 32 have an aneurysm before the bifurcation, which is a much less dangerous condition, or are healthy. The goal is to classify the patients using the morphology of their internal carotid artery. In this illustration we work with one of the observed variables, the radius. The data have previously been preprocessed, registered and smoothed and are observed on a grid of 2000 points in $[-100.3, 5.1]$, where the argument represents the distance between the observation point and the terminal bifurcation of the internal carotid artery, with positive values for points inside the skull. As we can see in Figure 3.3, the data are partially observed because the start and end points are different from subject to subject. All subjects are observed on the subset $\mathcal{I}_{\texttt{min}} = [-32.9, -7.4]$ that corresponds to 24.3 % of the whole domain.

We first apply the regularized linear classifiers to curves restricted to the common domain $\mathcal{I}_{\texttt{min}}$. The classification error estimated by cross-validation is 29.2 % for principal component, 29.2 % for conjugate gradient and 32.3 % for ridge regularized classification.

We compare this procedure with a different approach consisting of a multivariate classification method applied to principal component scores. Specifically, the covariance kernel is estimated from observations centred to their respective group means, its eigenfunctions are computed and quadratic discriminant analysis is applied to the inner products of the uncentred curves with the eigefunctions. This procedure is similar to that in Sangalli et al. (2009a). The best classifier of this type turns out to exhibit a misclassification error of 32.3 %, obtained with 2 eigenfunctions.

These values show that in this data set, when attention is restricted to the common domain $\mathcal{I}_{\texttt{min}}$, the proposed methodology is comparable to the more standard multivariate technique.

Next, we consider classification on extended domains including observed values outside the common domain $\mathcal{I}_{\texttt{min}}$. We build the sequence of domains $\mathcal{I}_{\texttt{min}}, \ldots, \mathcal{I}_{\texttt{max}}$ by enlarging the domain at each step by 1.25 % of the complement of $\mathcal{I}_{\texttt{min}}$. This step

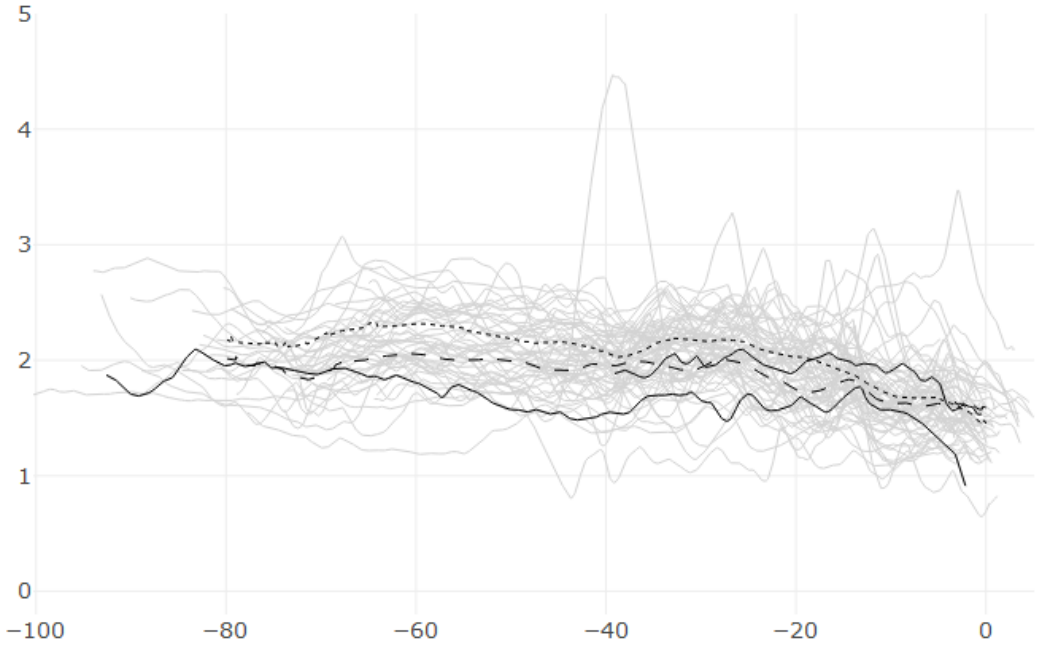

Figure 3.3: *Radius curves in the AneuRisk dataset, along with the mean for the group with an aneurysm after the bifurcation (dotted) and before the bifurcation or without aneurysm (dashed). Curves for two example subjects are highlighted by solid lines. Note the different start and end points for different subjects in the study.*

size is a compromise between the fineness of the grid and the computational cost. We consider extended domains up to $L = 40$, corresponding to $\mathcal{I}_{40} = [-66.6, -1.2]$, because not enough subjects have observed values outside this interval for reliable estimation and cross-validation. All regularized linear classification methods benefit from the domain extension, in particular the error rate for principal components drops from 29.2 % to 23.2 %, for conjugate gradients from 29.2 % to 25.8 % and for ridge regularization from 32.3 % to 25 %. The best domain is $\mathcal{I}_{10} = [-41.3, -5.8]$ for the conjugate gradient method and $\mathcal{I}_{11} = [-42.2, -5.7]$ for the other two methods.

The alternative method based on multivariate classification of scores cannot be applied on extended domains since the individual scores of incomplete curves cannot be computed, although they can be predicted (Kraus, 2015). By constrast, the proposed methods are entirely formulated in terms of distributional parameters, which can be consistently estimated from incomplete data, unlike individual quantites.

Appendix A

We consider three additional simulation studies, where we generate the data as in main simulation study, detailed in Section 2.4, with the only differences that

- in Simulation study 2 (Figures A.1 and A.2), the variance of the measurement error is increased to $\sigma_\varepsilon^2 = 0.4$;

- in Simulation study 3 (Figures A.3 and A.4), the variance of the spline coefficients is decreased to $\sigma^2 = 0.3$;

- in Simulation study 4 (Figures A.5 and A.6), the mean values of the spline coefficients of the first group of functional data are se to $\{\mu_{1,1}, \mu_{2,1}, \ldots, \mu_{20,1}\} = \{0, 0, 0, 0, 1, 2, 1, 0, -1, 1, 1.2, -1, 0, 0.5, 1, 0.5, 0, 0, 0, 0\}$, and the mean values of the spline coefficients of the second group are set to $\{\mu_{1,2}, \mu_{2,2}, \ldots, \mu_{20,2}\} = \{\mu_{20,1}, \mu_{19,1}, \ldots, \mu_{1,1}\}$.

Figures A.1, A.3 and A.5 show the data generated in the first replicates of these simulation studies. We implement the four techniques as detailed in Section 2.2. Figures A.2, A.4 and A.6 show the boxplots of the misclassification error and area under the ROC curve over the 50 simulation repetitions. Similar comments as those made for the main simulation study hold for all the considered simulation settings: by considering domain extensions it is possible to improve the discrimination results; on the other hand, even though a large part of the separation between the two groups lies outside of the common domain, considering the full domain where at least one of the data is observed leads to sub-optimal results.

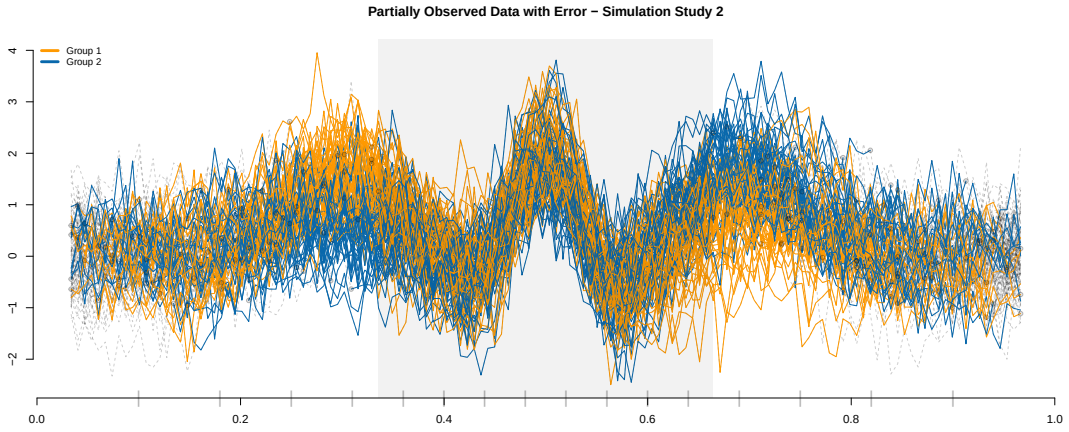

Figure A.1: *Simulation study 2. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

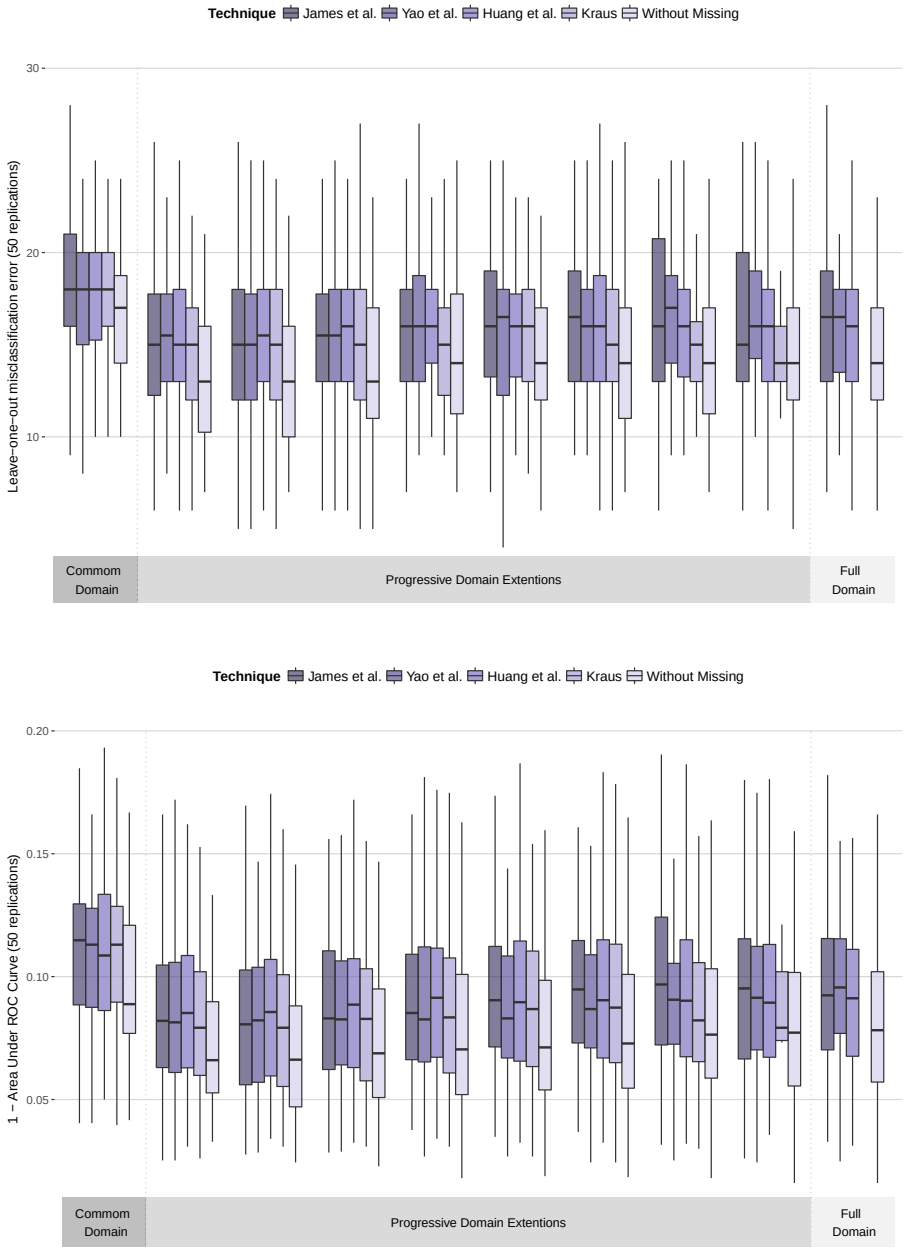

Figure A.2: *Simulation study 2. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

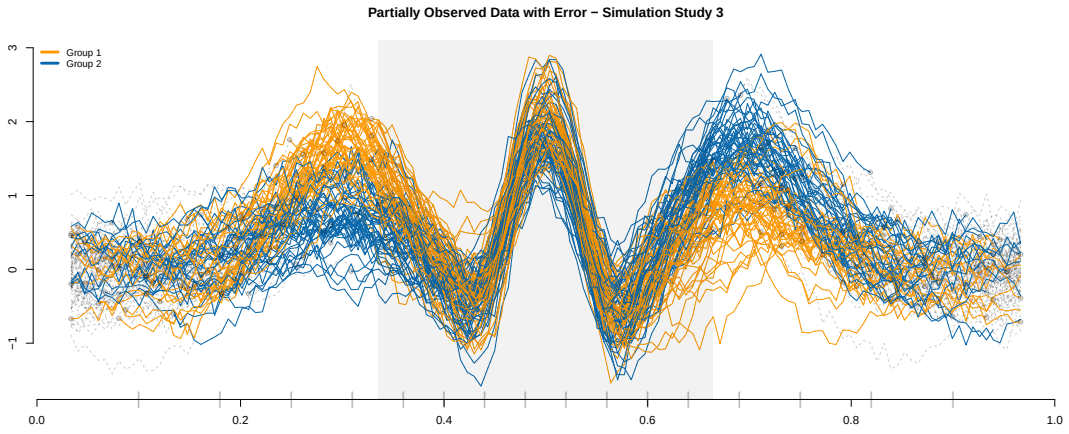

Figure A.3: *Simulation study 3. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

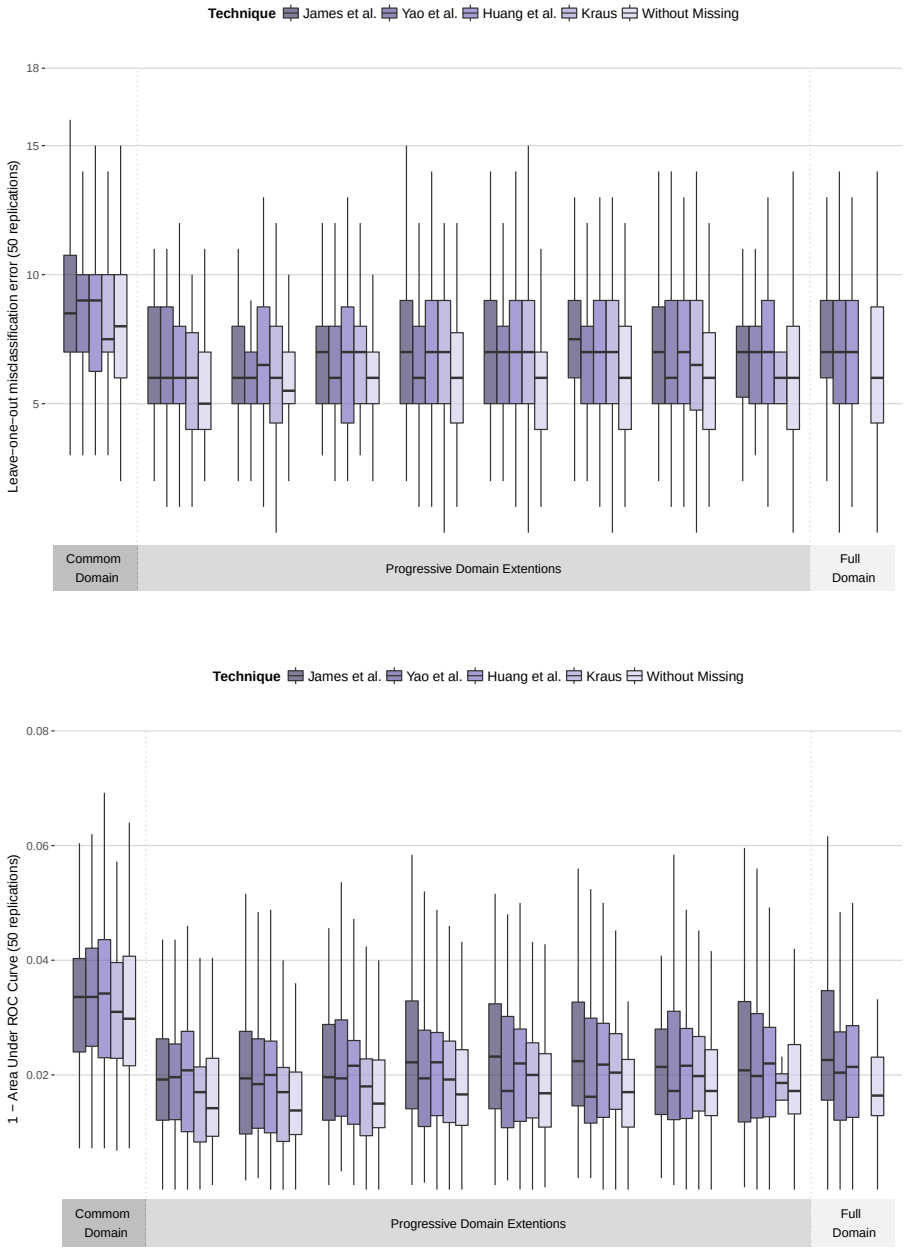

Figure A.4: *Simulation study 3. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

Figure A.5: *Simulation study 4. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

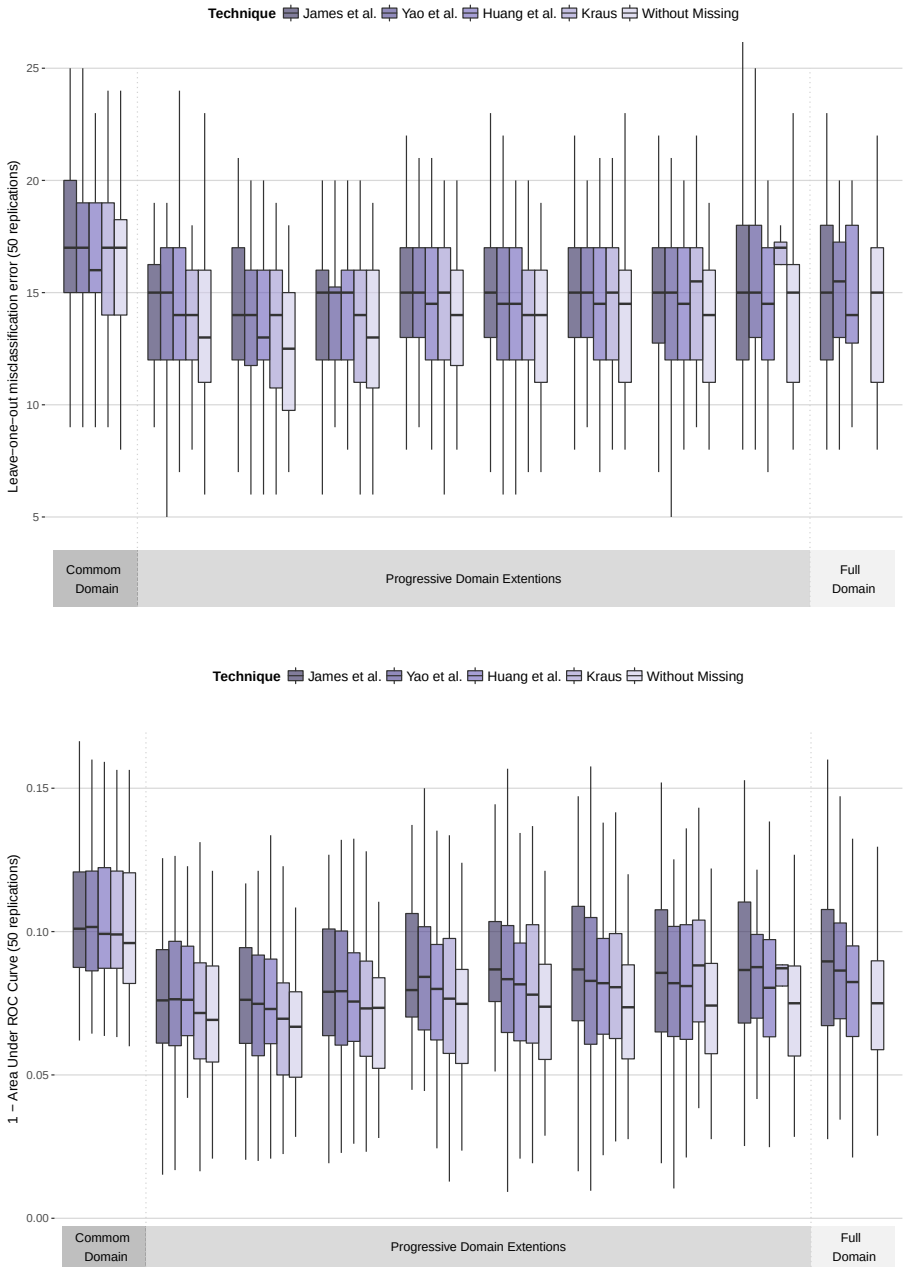

Figure A.6: *Simulation study 4. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

Appendix B

B.1 Proofs

B.1.1 Proof of Proposition 1

The misclassification probability for ψ_K is $D(\psi_K)$ given in (3.1). Since $\psi_K \in S_K$, we compute

$$\frac{|\langle \mu, \psi_K \rangle|}{\langle \psi_K, \mathscr{R}\psi_K \rangle^{1/2}} = \frac{\langle \mu, \mathscr{R}_K^- \mu \rangle}{\langle \mu, \mathscr{R}_K^- \mathscr{R} \mathscr{R}_K^- \mu \rangle^{1/2}} = \|(\mathscr{R}_K^-)^{1/2}\mu\|.$$

The right-hand side in the equation above converges by Lebesgue's monotone convergence theorem to $\|\mathscr{R}^{-1/2}\mu\|$, finite or infinite, and, therefore, the limiting misclassification probability that is attained along the regularization path ψ_K, $m \to \infty$ is $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

B.1.2 Proof of Proposition 2

The conjugate gradient method minimizes the quadratic objective function in the Krylov subspace $K_K(\mathscr{R}, \mu)$ whose elements are in the form $\eta = \sum_{k=0}^{K-1} c_k \mathscr{R}^k \mu = p(\mathscr{R})\mu$, where p is a polynomial of order lower than m. Then $\eta \in K_K(\mathscr{R}, \mu)$ can be written as $\eta = \sum_{j=1}^{\infty} p(\lambda_j) b_j \varphi_j$ with $b_j = \langle \mu, \varphi_j \rangle$. The objective function at η equals

$$\begin{aligned} \tfrac{1}{2}\langle \eta, \mathscr{R}\eta \rangle - \langle \mu, \eta \rangle &= \tfrac{1}{2}\langle p(\mathscr{R})\mu, \mathscr{R}p(\mathscr{R})\mu \rangle - \langle \mu, p(\mathscr{R})\mu \rangle \\ &= \sum_{j=1}^{\infty} b_j^2 \{ \tfrac{1}{2} p(\lambda_j)^2 \lambda_j - p(\lambda_j) \} \\ &= \sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_j)\{q(\lambda_j) - 2\}, \end{aligned} \tag{B.1}$$

where $q(\lambda) = p(\lambda)\lambda$ is a polynomial of degree at most m such that $q(0) = 0$. The conjugate gradient method finds the polynomial with these properties that minimizes the objective function. To prove the proposition we shall find a polynomial q with the required properties such that the objective function above is smaller than or equal to the objective function for the principal component classifier. The principal component classifier uses $\psi_K^{\mathtt{PC}} = \sum_{j=1}^{K} \lambda_j^{-1} b_j \varphi_j$, the objective function at $\psi_K^{\mathtt{PC}}$ equals

$$\tfrac{1}{2}\langle \psi_K^{\mathtt{PC}}, \mathscr{R}\psi_K^{\mathtt{PC}} \rangle - \langle \mu, \psi_K^{\mathtt{PC}} \rangle = -\sum_{j=1}^{K} \frac{b_j^2}{2\lambda_j}. \tag{B.2}$$

Consider the polynomial of degree m with $q(0) = 0$ given by

$$q(\lambda) = 1 - (-1)^K \frac{\lambda - \lambda_1}{\lambda_1} \cdots \frac{\lambda - \lambda_K}{\lambda_K}.$$

We see that $q(\lambda_j) = 1$ for $j = 1, \ldots, K$ and so the first K summands in the series (B.1) and (B.2) are equal. For $j > K$ it holds that $0 \leq q(\lambda_j) \leq 2$ due to the properties of the eigenvalue sequence, thus $q(\lambda_j)\{q(\lambda_j) - 2\} \leq 0$, and, therefore, the corresponding summands in the series (B.1) are negative whereas they are zero in the series (B.2). Hence for this polynomial

$$\sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_i)\{q(\lambda_i) - 2\} \leq -\sum_{j=1}^{K} \frac{b_j^2}{2\lambda_j}$$

and so the objective at the conjugate gradient solution must be smaller than or equal to the objective at the principal component solution. The inequality between the minima of the quadratic objective function implies the inequality between the misclassification probabilities stated in the proposition.

B.1.3 Proof of Proposition 3

Proceeding like in the proof of Proposition 1 we need to show that

$$\frac{\langle \mu, \mathscr{R}_\alpha^{-1}\mu \rangle}{\langle \mu, \mathscr{R}_\alpha^{-1}\mathscr{R}\mathscr{R}_\alpha^{-1}\mu \rangle^{1/2}} = \frac{\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j + \alpha}}{\left\{ \sum_{j=1}^{\infty} \frac{\lambda_j b_j^2}{(\lambda_j + \alpha)^2} \right\}^{1/2}} \xrightarrow[\alpha \to 0+]{} \left(\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j} \right)^{1/2} = \|\mathscr{R}^{-1/2}\mu\|,$$

where $b_j = \langle \mu, \varphi_j \rangle$ is the coefficient of μ in the eigenbasis. If $\sum_{j=1}^{\infty} b_j^2/\lambda_j < \infty$, the convergence follows from Lebesgue's monotone convergence theorem. Otherwise,

we use the inequality $\sum_{j=1}^{\infty} \lambda_j b_j^2/(\lambda_j+\alpha)^2 \leq \sum_{j=1}^{\infty} b_j^2/(\lambda_j+\alpha)$ to bound the expression on the left from below by $\{\sum_{j=1}^{\infty} b_j^2/(\lambda_j+\alpha)\}^{1/2}$ which diverges to infinity again by Lebesgue's theorem.

B.1.4 Proof of Theorem 1

The probability of misclassifying a new observation using the conjugate gradient classifier based on $\widehat{\psi}_{K_N}^{\mathrm{CG}}$ is $D(\widehat{\psi}_{K_N}^{\mathrm{CG}}) = 1-\Phi(|Q(\widehat{\psi}_{K_N}^{\mathrm{CG}})|/2)$. We need to show that the fraction in $Q(\widehat{\psi}_{K_N}^{\mathrm{CG}})$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ along the regularization path satisfying the assumptions of the theorem. To deal with the numerator in $Q(\widehat{\psi}_{K_N}^{\mathrm{CG}})$ we can show that

$$\langle \mu, \widehat{\psi}_{K_N}^{\mathrm{CG}} \rangle - \langle \mu, \psi_{K_N}^{\mathrm{CG}} \rangle = O_P(N^{-1/2}\omega_{K_N}^{-1}\|\gamma^{(m_N)}\| + N^{-1}\omega_{K_N}^{-3}). \tag{B.3}$$

This result follows from an analog of (5.9) in Theorem 5.3 in Delaigle and Hall (2012b) and intermediate results in the proof of that theorem which can be established in our context. Specifically, the necessary modifications of the proofs of Theorems 5.1, 5.2 and 5.3 in Delaigle and Hall (2012b) are as follows. All results remain valid for incomplete instead of complete curves because the proofs depend only on the root-n consistency of the covariance estimators which is satisfied for functional fragments (Kraus, 2015, Proposition 1). Moreover, derivations in Delaigle and Hall (2012b) can be repeated without assuming that the theoretical solution $\psi = \mathscr{R}^{-1}\mu$ exists as an element of $L^2(\mathcal{I})$. Indeed, the proofs in Delaigle and Hall (2012b) are based on stochastic expansions of $\widehat{\mathscr{R}}^j\psi = \widehat{\mathscr{R}}^j\mathscr{R}^{-1}\mu$, in our notation, around $\mathscr{R}^j\psi = \mathscr{R}^j\mathscr{R}^{-1}\mu = \mathscr{R}^{j-1}\mu$ and derived quantities but the same steps can be done for $\widehat{\mathscr{R}}^{j-1}\widehat{\mu}$ around $\mathscr{R}^{j-1}\mu$ present in our situation. In other words, it holds that $\widehat{\psi}_{K_N}^{\mathrm{CG}}$ and $\psi_{K_N}^{\mathrm{CG}}$ converge to each other without assuming that $\psi_{K_N}^{\mathrm{CG}}$ converges. Similarly, for the denominator in $Q(\widehat{\psi}_{K_N}^{\mathrm{CG}})$ it holds that

$$\langle \widehat{\psi}_{K_N}^{\mathrm{CG}}, \mathscr{R}\widehat{\psi}_{K_N}^{\mathrm{CG}} \rangle - \langle \psi_{K_N}^{\mathrm{CG}}, \mathscr{R}\psi_{K_N}^{\mathrm{CG}} \rangle = O_P(N^{-1/2}\omega_{K_N}^{-1}\|\gamma^{(m_N)}\| + N^{-1}\omega_{K_N}^{-3}). \tag{B.4}$$

This last result is analogous to (7.27) of Delaigle and Hall (2012b) whose proof can be repeated with the same modifications for our situation as before. Therefore, regardless of whether $\|\mathscr{R}^{-1}\mu\|$ or $\|\mathscr{R}^{-1/2}\mu\|$ is finite or infinite, we see that the theoretical and empirical regularized quantities approach each other at rates

given in (B.3) and (B.4). The result on $D(\widehat{\psi}^{\mathrm{CG}}_{K_N})$ then follows like in the proof of Proposition 1.

B.1.5 Proof of Theorem 2

We show that $D(\widehat{\psi}^{\mathtt{PC}}_{K_N}) = 1 - \Phi(|Q(\widehat{\psi}^{\mathtt{PC}}_{K_N})|/2)$ converges in probability to $1 - \Phi(\|\mathscr{R}^{-1/2}\|/2)$. The strategy of the proof is similar to that for Theorem 3.1 of Cardot et al. (1999) for the principal component approach to the functional linear model. The difference is in the incompleteness of the functional data and in that we do not assume that the underlying theoretical inverse problem has a solution. We rewrite

$$\|\widehat{\psi}^{\mathtt{PC}}_{K_N} - \psi^{\mathtt{PC}}_{K_N}\| \leq \|\widehat{\mathscr{R}}^{-}_{K_N} - \mathscr{R}^{-}_{K_N}\|_\infty \|\widehat{\mu}\| + \|\mathscr{R}^{-}_{K_N}\|_\infty \|\widehat{\mu} - \mu\|.$$

Proceeding like in the proof of Lemma 5.1 in Cardot et al. (1999) we can show that

$$\|\widehat{\mathscr{R}}^{-}_{K_N} - \mathscr{R}^{-}_{K_N}\|_\infty \leq \widehat{\lambda}^{-1}_{K_N} \lambda^{-1}_{K_N} \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty + 2\lambda^{-1}_{K_N} \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty \sum_{j=1}^{K_N} a_j.$$

Here $\widehat{\lambda}_j$ are the eigenvalues of $\widehat{\mathscr{R}}$ in descending order and $\widehat{\varphi}_j$ are the corresponding eigenfunctions. When establishing the above inequality one uses the facts that $|\widehat{\lambda}_j - \lambda_j| \leq \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty$ and $\|\widehat{\varphi}_j - \mathrm{sign}\langle\widehat{\varphi}_j, \varphi_j\rangle\varphi_j\| \leq a_j \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty$ which are known from (Bosq, 2000, Lemmas 4.2 and 4.3) for the empirical covariance operator from complete curves but hold also for functional fragments, see the proof of Proposition 2 in the supplementary document for Kraus (2015). Since $\|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty = O_P(N^{-1/2})$, we see that $\widehat{\lambda}^{-1}_{K_N} \lambda^{-1}_{K_N} \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty 1_{[\widehat{\lambda}_{K_N} > \lambda_{K_N}/2]} \leq 2\lambda^{-2}_{K_N} \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{K_N} O_P(N^{-1/2})$. Since the probability of the event $[\widehat{\lambda}_{K_N} < \lambda_{K_N}/2]$ is bounded by $\lambda^{-2}_{K_N} O(N^{-1})$, hence converges to 0, it follows that $\widehat{\lambda}^{-1}_{K_N} \lambda^{-1}_{K_N} \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{K_N} O_P(N^{-1/2})$. Combining this with the facts that $\|\widehat{\mu}\| = O_P(1)$, $\|\mathscr{R}^{-}_{K_N}\| = \lambda^{-1}_{K_N}$ and $\|\widehat{\mu} - \mu\| = O_P(N^{-1/2})$ gives that

$$\|\widehat{\psi}^{\mathtt{PC}}_{K_N} - \psi^{\mathtt{PC}}_{K_N}\| \leq \lambda^{-2}_{K_N} O_P(N^{-1/2}) + \lambda^{-1}_{K_N} O_P(N^{-1/2}) \sum_{j=1}^{K_N} a_j.$$

Similar arguments can be used in the analysis of the denominator in $Q(\widehat{\psi}^{\mathtt{PC}}_{K_N})$. In conclusion, we obtain that the estimation errors for the quantities in the numerator and denominator converge to 0 at rates

$$\langle \mu, \widehat{\psi}^{\mathtt{PC}}_{K_N} \rangle - \langle \mu, \psi^{\mathtt{PC}}_{K_N} \rangle = \lambda^{-2}_{K_N} O_P(N^{-1/2}) + \lambda^{-1}_{K_N} O_P(N^{-1/2}) \sum_{j=1}^{K_N} a_j, \quad \text{(B.5)}$$

$$\langle \widehat{\psi}^{\mathtt{PC}}_{K_N}, \mathscr{R}\widehat{\psi}^{\mathtt{PC}}_{K_N} \rangle - \langle \psi^{\mathtt{PC}}_{K_N}, \mathscr{R}\psi^{\mathtt{PC}}_{K_N} \rangle = \lambda^{-2}_{K_N} O_P(N^{-1/2}) + \lambda^{-1}_{K_N} O_P(N^{-1/2}) \sum_{j=1}^{K_N} a_j. \quad \text{(B.6)}$$

In light of (B.5) and (B.6) the asymptotic behaviour of the misclassification probability is driven by the behaviour of the theoretical classifier addressed in Proposition 1.

B.1.6 Proof of Theorem 3

We show that the fraction $|Q(\widehat{\psi}^{\mathtt{R}}_{K_N})|$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ as $N \to \infty$. For the numerator we write

$$\langle \mu, \widehat{\psi}^{\mathtt{R}}_{\alpha_N} \rangle - \langle \mu, \mathscr{R}^{-1}_{\alpha_N}\mu \rangle = \langle \mu, (\widehat{\mathscr{R}}^{-1}_{\alpha_N} - \mathscr{R}^{-1}_{\alpha_N})\widehat{\mu} \rangle + \langle \mu, \mathscr{R}^{-1}_{\alpha_N}(\widehat{\mu} - \mu) \rangle. \quad \text{(B.7)}$$

For the first term on the right we compute

$$\begin{aligned} |\langle \mu, (\widehat{\mathscr{R}}^{-1}_{\alpha_N} - \mathscr{R}^{-1}_{\alpha_N})\widehat{\mu} \rangle| &\leq \|\mu\| \|\widehat{\mathscr{R}}^{-1}_{\alpha_N} - \mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mu}\| \\ &= \|\mu\| \|\widehat{\mathscr{R}}^{-1}_{\alpha_N}(\widehat{\mathscr{R}}_{\alpha_N} - \mathscr{R}_{\alpha_N})\mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mu}\| \\ &\leq \|\mu\| \|\widehat{\mathscr{R}}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mathscr{R}}_{\alpha_N} - \mathscr{R}_{\alpha_N}\|_\infty \|\mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mu}\| \\ &\leq \alpha_N^{-2} O_P(N^{-1/2}), \end{aligned}$$

since $\|\widehat{\mathscr{R}}^{-1}_{\alpha_N}\|_\infty \leq \alpha_N^{-1}$, $\|\mathscr{R}^{-1}_{\alpha_N}\|_\infty \leq \alpha_N^{-1}$, $\|\widehat{\mu}\| = O_P(1)$ and $\|\widehat{\mathscr{R}}_{\alpha_N} - \mathscr{R}_{\alpha_N}\|_\infty = \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty = O_P((N_0 + N_1)^{-1/2})$ (Kraus, 2015, Proposition 1). For the second term on the right side of (B.7) we obtain

$$|\langle \mu, \mathscr{R}^{-1}_{\alpha_N}(\widehat{\mu} - \mu) \rangle| \leq \|\mu\| \|\mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mu} - \mu\| \leq \alpha_N^{-1} O_P(N^{-1/2}).$$

The quantity in the denominator in $Q(\widehat{\psi}^{\mathtt{R}}_{K_N})$ can be rewritten as

$$\langle \widehat{\psi}^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\widehat{\psi}^{\mathtt{R}}_{\alpha_N} \rangle - \langle \psi^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\psi^{\mathtt{R}}_{\alpha_N} \rangle = \langle \widehat{\psi}^{\mathtt{R}}_{\alpha_N} - \psi^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\widehat{\psi}^{\mathtt{R}}_{\alpha_N} \rangle + \langle \psi^{\mathtt{R}}_{\alpha_N}, \mathscr{R}(\widehat{\psi}^{\mathtt{R}}_{\alpha_N} - \psi^{\mathtt{R}}_{\alpha_N}) \rangle. \quad \text{(B.8)}$$

The first term on the right is

$$\begin{aligned}\langle \widehat{\psi}^{\mathtt{R}}_{\alpha_N} - \psi^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\widehat{\psi}^{\mathtt{R}}_{\alpha_N}\rangle &= \langle \widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu} - \mathscr{R}^{-1}_{\alpha_N}\mu, \mathscr{R}\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}\rangle \\ &= \langle \mathscr{R}^{-1}_{\alpha_N}(\mathscr{R}_{\alpha_N} - \widehat{\mathscr{R}}_{\alpha_N})\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}, \mathscr{R}\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}\rangle + \langle \mathscr{R}^{-1}_{\alpha_N}(\widehat{\mu}-\mu), \mathscr{R}\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}\rangle.\end{aligned}$$

Here we compute for the first summand

$$\begin{aligned}|\langle \mathscr{R}^{-1}_{\alpha_N}(\mathscr{R}_{\alpha_N} - \widehat{\mathscr{R}}_{\alpha_N})\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}, \mathscr{R}\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}\rangle| &\leq \|\widehat{\mu}\|^2 \|\widehat{\mathscr{R}}^{-1}_{\alpha_N}\|^2_\infty \|\mathscr{R}\mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mathscr{R}} - \mathscr{R}\|_\infty \\ &\leq \alpha_N^{-2} O_P(N^{-1/2})\end{aligned}$$

using properties mentioned previously and $\|\mathscr{R}\mathscr{R}^{-1}_{\alpha_N}\|_\infty \leq 1$ and for the second summand

$$|\langle \mathscr{R}^{-1}_{\alpha_N}(\widehat{\mu}-\mu), \mathscr{R}\widehat{\mathscr{R}}^{-1}_{\alpha_N}\widehat{\mu}\rangle| \leq \|\mathscr{R}\mathscr{R}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mathscr{R}}^{-1}_{\alpha_N}\|_\infty \|\widehat{\mu}-\mu\| \leq \alpha_N^{-1} O_P(N^{-1/2}).$$

Putting these results together we see that the absolute value of the first term on the right side in (B.8) is dominated by $\alpha_N^{-2} O_P(N^{-1/2})$. The second term on the right in (B.8) can be analyzed like the first two terms on the right in (B.7) with $\mathscr{R}\mathscr{R}^{-1}_{\alpha_N}\mu$ in place of μ. This way we bound its absolute value from above by $\alpha_N^{-2} O_P(N^{-1/2})$. These results imply that the estimation errors vanish at rates given by

$$\langle \mu, \widehat{\psi}^{\mathtt{R}}_{\alpha_N}\rangle - \langle \mu, \psi^{\mathtt{R}}_{\alpha_N}\rangle = \alpha_N^{-2} O_P(N^{-1/2}), \tag{B.9}$$

$$\langle \widehat{\psi}^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\widehat{\psi}^{\mathtt{R}}_{\alpha_N}\rangle - \langle \psi^{\mathtt{R}}_{\alpha_N}, \mathscr{R}\psi^{\mathtt{R}}_{\alpha_N}\rangle = \alpha_N^{-2} O_P(N^{-1/2}). \tag{B.10}$$

Hence the empirical classifier has the same limiting error as the theoretical one addressed in Proposition 3.

Bibliography

Baíllo, A., Cuevas, A. and Cuesta-Albertos, J. A. (2011a) Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics*, **38**, 480–498.

Baíllo, A., Cuevas, A. and Fraiman, R. (2011b) Classification methods for functional data. In *The Oxford Handbook of Functional Data Analysis*, 259–297. Oxford University Press, Oxford.

Berlinet, A., Biau, G. and Rouviere, L. (2008) Functional supervised classification with wavelets. *Annales de l'ISUP*, **52**, 19.

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2016) Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, **26**, 619–638.

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2017) On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association*. To appear.

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

Blanchard, G. and Krämer, N. (2010) Kernel partial least squares is universally consistent. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 57–64.

Bongiorno, E. G. and Goia, A. (2016) Classification methods for hilbert data based on surrogate density. *Computational Statistics & Data Analysis*, **99**, 204–222.

Bosq, D. (2000) *Linear Processes in Function Spaces*. New York: Springer.

Bugni, F. A. (2012) Specification test for missing functional data. *Econometric Theory*, **28**, 959–1002.

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

Cardot, H., Mas, A. and Sarda, P. (2007) CLT in functional linear regression models. *Probability Theory and Related Fields*, **138**, 325–361.

Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R. and Matrán, C. (2007) The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, **51**, 4814–4831.

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

Dai, X., Hadjipantelis, P. Z., Ji, H., Müller, H.-G. and Wang, J.-L. (2017a) *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.3.0.

Dai, X., Müller, H.-G. and Yao, F. (2017b) Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, **104**, 545–560.

De Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–263.

Delaigle, A. and Hall, P. (2012a) Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **74**, 267–286.

— (2012b) Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, **40**, 322–352.

— (2013) Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.

— (2016) Approximating fragmented functional data by segments of markov chains. *Biometrika*, **103**, 779–799.

Delaigle, A., Hall, P. and Bathia, N. (2012) Componentwise classification and clustering of functional data. *Biometrika*, **99**, 299–313.

Di, C.-Z., Crainiceanu, C. M. and Jank, W. S. (2014) Multilevel sparse functional principal component analysis. *Stat*, **3**, 126–143.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2007) Functional pls logit regression model. *Comput. Stat. Data Anal.*, **51**, 4891–4902.

Febrero-Bande, M., Galeano, P. and González-Manteiga, W. (2017) Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, **85**, 61–83.

Ferraty, F., Hall, P. and Vieu, P. (2010) Most-predictive design points for functional data predictors. *Biometrika*, **97**, 807–824.

Ferraty, F. and Vieu, P. (2003) Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, **44**, 161 – 173.

Floriello, D. and Vitelli, V. (2017) Sparse clustering of functional data. *Journal of Multivariate Analysis*, **154**, 1 – 18.

Glendinning, R. H. and Herbert, R. A. (2003) Shape classification using smooth principal components. *Pattern Recogn. Lett.*, **24**, 2021–2030.

Goldberg, Y., Ritov, Y. and Mandelbaum, A. (2014) Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference*, **147**, 53 – 65.

Gromenko, O., Kokoszka, P. and Sojka, J. (2017) Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Stat.*, **11**, 898–918.

Hall, P., Poskitt, D. S. and Presnell, B. (2001) A functional data–analytic approach to signal discrimination. *Technometrics*, **43**, 1–9.

Hastie, T. J., Tibshirani, R. and Friedman, J. H. (2009) *The Elements of Statistical Learning*. Springer, New York.

Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*. New York: Springer.

Hsing, T. and Eubank, R. (2015) *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Huang, J. Z., Shen, H. and Buja, A. (2008) Functional principal components analysis via penalized rank one approximation. *Electron. J. Statist.*, **2**, 678–695.

Izenman, A. J. (2009) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer Science & Business Media.

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

James, G. M., Hastie, T. J. and Sugar, C. A. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

Kokoszka, P. and Reimherr, M. (2017) *Introduction to Functional Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC.

Kraus, D. (2015) Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 777–801.

Leng, X. and Müller, H.-G. (2006) Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68–76.

Liebl, D. (2013) Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.*, **7**, 1562–1592.

Liebl, D. and Rameseder, S. (2018) Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis*.

Lila, E., Aston, J. A. D. and Sangalli, L. M. (2016) Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.*, **10**, 1854–1879.

Lingjærde, O. C. and Christophersen, N. (2000) Shrinkage structure of partial least squares. *Scandinavian Journal of Statistics. Theory and Applications*, **27**, 459–473.

Liu, C., Ray, S. and Hooker, G. (2017) Functional principal component analysis of spatially correlated data. *Statistics and Computing*, **27**, 1639–1654.

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

Peng, J. and Paul, D. (2011) *fpca: Restricted MLE for Functional Principal Components Analysis*. R package version 0.2-1.

Phatak, A. and de Hoog, F. (2002) Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, **16**, 361–367.

Pini, A. and Vantini, S. (2016) The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics*, **72**, 835–845.

— (2017) Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, **29**, 407–424.

Pini, A., Vantini, S., Colosimo, B. M. and Grasso, M. (2017) Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Preda, C., Saporta, G. and Lévéder, C. (2007) Pls classification of functional data. *Computational Statistics*, **22**, 223–235.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer Series in Statistics. Springer.

Rossi, F. and Villa, N. (2006) Support vector machine for functional data classification. *Neurocomput.*, **69**, 730–742.

Sangalli, L. M., Secchi, P. and Vantini, S. (2014a) Analysis of aneurisk65 data: k-mean alignment. *Electron. J. Statist.*, **8**, 1891–1904.

— (2014b) Aneurisk65: A dataset of three-dimensional cerebral vascular geometries. *Electron. J. Statist.*, **8**, 1879–1890.

Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009a) A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, **104**, 37–48.

— (2009b) Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**, 285–306.

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010) k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, **54**, 1219 – 1233.

Shin, H. (2008) An extension of fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, **99**, 1191 – 1216.

Song, J. J., Deng, W., Lee, H.-J. and Kwon, D. (2008) Optimal classification for time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*, **32**, 426 – 432.

Stefanucci, M., Sangalli, L. M. and Brutti, P. (2018) PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. *Statistica Neerlandica*. To appear.

Wang, X., Ray, S. and Mallick, B. K. (2007) Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, **102**, 962–973.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.